Hayit Greenspan
Henning Müller
Tanveer Syeda-Mahmood (Eds.)

LNCS 7723

Medical Content-Based Retrieval for Clinical Decision Support

Third MICCAI International Workshop, MCBR-CDS 2012
Nice, France, October 2012
Revised Selected Papers

Springer

Lecture Notes in Computer Science 7723

Hayit Greenspan Henning Müller
Tanveer Syeda-Mahmood (Eds.)

Medical Content-Based Retrieval for Clinical Decision Support

Third MICCAI International Workshop, MCBR-CDS 2012
Nice, France, October 1, 2012
Revised Selected Papers

Springer

Volume Editors

Hayit Greenspan
Tel Aviv University
The Iby and Aladar Fleischmann Faculty of Engineering
Ramat Aviv, Israel
E-mail: hayit@eng.tau.ac.il

Henning Müller
University of Applied Sciences Western Switzerland (HES-SO)
Business Information Systems
TechnoArk 3, 3960 Sierre, Switzerland
E-mail: henning.mueller@hevs.ch

Tanveer Syeda-Mahmood
IBM Almaden Research Center
Multi-modal Mining for Healthcare
650 Harry Road
San Jose, CA 95120, USA
E-mail: stf@almaden.ibm.com

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36677-2 e-ISBN 978-3-642-36678-9
DOI 10.1007/978-3-642-36678-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013931221

CR Subject Classification (1998): J.3, I.5, H.2.8, I.4, H.3-5, I.2

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This document contains articles from the Third Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MCBR-CDS) that took place at the MICCAI (Medical Image Computing for Computer Assisted Intervention) 2012 conference in Nice, France, on October 1, 2012. The first workshop on this topic took place at MICCAI 2009 in London, UK. The second one was held in Toronto, Canada, in 2011. An earlier workshop on medical image retrieval was conducted at MICCAI 2007, in Brisbane, Australia.

The workshop obtained 15 high-quality submissions of which 10 were selected for presentation. Three external reviewers and one workshop organizer reviewed each of the papers. The review process was double blind.

In addition to the oral presentations, two invited presentations were given. Gwenole Quellec, the first invited speaker, presented a talk on heterogeneous information retrieval from medical databases. His talk developed several ideas about medical image and video retrieval applications from a theoretical and an application perspective. Georg Langs presented his invited presentation on the VISCERAL project that aims at creating a benchmark for medical imaging on extremely large data sets using a cloud–based infrastructure that is shared by the participants. A panel at the end discussed the role of content-based image retrieval in clinical decision support. In general, the workshop resulted in many lively discussions and showed well the current trends and tendencies in content-based medical retrieval and how this can support decisions in clinical work.

These proceedings contain the ten accepted papers of the workshop as well as the two invited presentations. An overview of the workshop initiates the proceedings, summarizing the papers and the discussions that took place at the workshop itself.

We would like to thank all the reviewers that helped make a selection of high-quality papers for the workshop. We hope to have a similar workshop at next year's MICCAI conference.

December 2012

Hayit Greenspan
Henning Müller
Tanveer Syeda-Mahmood

Organization

General Co-chairs	Hayit Greenspan, Israel Henning Müller, Switzerland Tanveer Syeda-Mahmood, USA
Publication Chair	Hayit Greenspan, Israel

International Program Committee

Burak Acar	Bogazici University, Turkey
Sameer Anatani	National Library of Medicine (NLM), USA
Rahul Bhotika	GE Global Research Center, NY, USA
Albert Chung	Hong Kong University of Science and Technology, Hong Kong
Adrien Depeursinge	University of Applied Sciences Western Switzerland (HES-SO), Switzerland
Thomas M. Deserno	Aachen University of Technology (RWTH), Germany
Gerhard Engelbrecht	University Pompeu Fabra (UPF), Spain
Bram van Ginneken	Radboud University Nijmegen Medical Centre, The Netherlands
Allan Hanbury	Vienna University of Technology, Austria
Jayashree Kalpathy-Cramer	Harvard University, USA
Nico Karssemeijer	Radboud University Nijmegen, The Netherlands
Rodney Long	National Library of Medicine, USA
Kazunori Okada	San Francisco State University, USA
Daniel Racoceanu	French National Center for Scientific Research (CNRS), France
Daniel Rubin	Stanford University, USA
Linda Shapiro	University of Washington, US
Ron Summers	National Institutes of Health, USA
Agma Traina	University of Sao Paulo, Brazil
S. Kevin Zhou	Siemens Corporate Research, USA

Sponsors

European Commission 7th Framework Programme, project Khresmoi (257528)

Table of Contents

Workshop Overview

Invited Talk

Methods

3D/4D Data Retrieval

Invited Talk

Visual Features

Multimodal Retrieval

Overview of the Third Workshop on Medical Content–Based Retrieval for Clinical Decision Support (MCBR–CDS 2012)

Henning Müller[1,2] and Hayit Greenspan[3]

[1] University of Applied Sciences Western Switzerland (HES–SO), Switzerland
[2] University Hospitals and University of Geneva, Switzerland
[3] University of Tel Aviv, Israel
henning.mueller@hevs.ch

Abstract. The third workshop on Medical Content–based Retrieval for Clinical Decision Support (MCBR–CDS 2012) took place in connection with the MICCAI conference (Medical Image Computing for Computer–Assisted Intervention) in Nice, France on October 1, 2012. This text gives an overview of the invited presentations and scientific papers presented at the workshop. In the description of the papers the comments and discussions at the workshop are taken into account, highlighting the current tendencies and scientific merits. The workshop finished with a panel that discussed the need of clients of image retrieval software and additional important areas such as the importance of high–quality annotated training and test data sets to advance current research. Such big data sets and a framework for researchers to work on them can have an important impact on the field of image–based decision support in the future.

Keywords: medical image analysis, medical information retrieval, clinical decision support, content–based medical image retrieval.

1 Introduction

Visual image retrieval or content–based image retrieval (CBIR) has started in the early 1990s as an increasing amount of image data had become available in digital form and could thus be analyzed visually through computers [10]. In the following decade much research work happened in the field and image retrieval advanced strongly in terms of architectures and also visual features with many application domains [19].

In the medical field, propositions for image retrieval and its usefulness were made quite early [12,21]. Still, clinical applications were very rare despite a large amount of research in the field as stated a review article in 2004 [14]. Since then, medical image retrieval research has exploded as data sets have become available in benchmarks [13,8]. A more recent review article highlights these developments and further current challenges in medical image retrieval [2].

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 1–9, 2013.

Medical imaging is producing very large amounts of data and although many articles discuss retrieval on an entire PACS (Picture Archival and Communication System) [5], no current system has indexed such large amounts. A report of the European Union estimates that 30% of world storage was occupied in 2010 by medical imaging and that mammographies in the USA alone accounted for 2.5 Petabytes in 2009 [1]. The influence of the analysis of big data can be major for medical image retrieval as the scalability to work with extremely large data sets could enable researchers to tackle rare diseases and really increase diagnosis performance based on learning from existing data.

The workshop Medical Content–Based Retrieval for Clinical Decision Support (MCBR–CDS) was held for the third time in connection with MICCAI in 2012. The workshop received 15 high–quality submissions and also asked two invited speakers to submit a paper to these proceedings. All papers were reviewed by at least three independent reviewers and in addition by one workshop organizer. Based on the review results ten papers were finally accepted for the workshop in addition to the two invited papers. These papers are published in the workshop proceedings of this volume of the Springer Lecture Notes in Computer Science. A panel finalized the workshop program. This panel led to vivid discussions on the role of medical image retrieval in clinical practice and also the sharing of annotated medical data to advance research towards large–scale or *big data*.

Authors were able to modify their paper until two weeks after the workshop based on the comments received during the workshop and based on the discussions that took place during the entire day. The workshop presentations included several of the important current research areas including the increasing analysis of multidimensional data, the road towards the use of big data, but also the important topics of using high quality visual features and building real applications based on existing techniques including combinations of text and visual analysis. The workshop attendance reflected the fact that this is an important topic for the MICCAI community.

This paper starts with an overview of the papers presented at the workshop, starting with the invited talks in Section 2. Section 3 summarizes the discussions that took place at the panel session and throughout the workshop, and Section 4 closes the paper with conclusions.

2 Papers Presented at the Workshop

This Section describes the two invited talks of the workshop and also all scientific presentations.

2.1 Invited Presentations

Gwenole Quellec was the first invited speaker starting the workshop with a talk on *heterogeneous information retrieval from medical databases*. His inspiring talk developed several ideas about medical image and video retrieval applications from a theoretical but also from an application perspective. The importance of clinical data in connection with the visual features was highlighted for case–based retrieval, that can be considered much closer to clinical routine than image–based retrieval. Several applications for case–based retrieval were analyzed, notably the use of retinopathy images. A second part of the presentation analyzed challenges for retrieval of videos of medical interventions such as cataract operations as an example. The paper published in these proceedings [17] describes only part of the presentation, notably the video retrieval part. For the video analysis several videos of surgical acts of young and experienced surgeons were analyzed. Specific phases of each operation were found and could then be detected automatically in the videos and separated. Inside each phase a real–time analysis is performed to be able to react quickly to deviations from an optimal operation. Operations considered good and operations considered poor are compared and thus for an on–going operation any deviation is detected in real time so the surgeon can be informed about a potential risk. A particular value of such a technique would be in the training of young surgeons and in giving constant feedback in real operations and warning from potential dangers..

Georg Langs presented his invited presentation on *VISCERAL: towards large data in medical imaging — challenges and directions* in the afternoon. This presentation explained the VISCERAL[1] project that aims at creating a benchmark for medical imaging on extremely large data sets using a cloud–based infrastructure that is shared by the participants [11]. The goals of the benchmark include two challenges for the research community on a data set of at least 10–20 TB of medical image data that are available to the project. The first challenge is the identification of organs or reference points in the human body, whether in full body scans or partial volumetric data. A focus of the project will thus clearly be on 3D data. A second comparison aims at retrieval using the radiology reports in potentially different languages and with various types of images to find similar cases. The discussion that followed the presentation showed the interest in the topic and highlighted that there are still many things that need to be defined by the project in collaboration with the research community. Work at the medical University of Vienna and inside the Khresmoi[2] project were then presented showing how important an efficient and effective data analysis is when going towards big data for a good retrieval quality. The detection of regions of interest or at least the identification of organs in the body can be an important first step to better analyze the visual medical data and extract semantic labels from the visual image information automatically.

2.2 3D Methods

3D data analysis has grown substantially in medical image retrieval over the past years and this was shown by half of the submissions to the workshop dealing with retrieval of tomographic data sets. It is the quickest growing medical data type.

[1] `http://visceral.eu/`, VISual Concept Extraction challenge in RAdioLogy

[2] `http://www.khresmoi.eu/`

In [4], Catalano at al. describe their work titled *exploiting 3D part–based analysis, description and indexing to support medical applications*. The text highlights the importance of the retrieval of medical 3D data, notably surface–based models for analyzing the different parts of objects and similarities between these parts. Modeling of medical 3D data is an important topic and then being able to use this information for similarity–based retrieval is equally important for many applications.

Another 3D retrieval application was presented by Indriyati Atmosukarto with *skull retrieval for craniosynostosis using sparse logistic regression models* [23]. This approach deals with malformations of the head bones regarding craniosynostosis. The 3D analysis of the skull can help to identify and more importantly quantify certain malformations. The analysis following operative interventions can help to track the evolution of the skull over time. For the interesting application and the solid theoretical and methodological quality this paper was awarded with the Khresmoi prize of the best workshop paper.

2.3 3D/4D Retrieval

Besides the general volume–based 3D analysis and applications, there are also several applications with clear medical application scenarios.

In *retrieval of 4D dual energy CT for pulmonary embolism diagnosis* Foncubierta et al. describe an application of using the 4D data of dual energy CT for the detection of pulmonary embolism in emergency radiology [6]. A difficulty is the extremely large amount of data that needs to be analyzed (11 times 400 slices) and also the difficulty to find in which energy bands the discriminative information is contained. As solid 4D texture is concerned it is also extremely difficult to visualize the data sets.

Simonyan et al. describe in *immediate ROI search for 3–D medical images* [18] a retrieval system in 3D databases that allows users to select 3D regions of interest and then search for visually similar volumes of interest in other image series. The paper gives examples for a theoretically sound framework using the ADNI (Alzheimer disease neuro imaging initiative) database of MRI images to demonstrate the system experimentally.

In *synergy of 3D SIFT and sparse codes for classification of viewpoints from echocardiogram videos* Qian et al. describe the analysis of echocardiogram 3D data [15]. 3D SIFT features are used on the noisy data sets of children ultrasound data of the heart. Modern ultrasound really allows to have high quality data where automatic analysis can become possible. Even 4D data sets, for example 3D data of the beating heart, have become available and could be intersting for similarity–based retrieval applications.

Quatrehomme et al. present in *assessing the classification of liver focal lesions by using multi–phase computer tomography scans* an interesting approach to the analysis of liver lesions [16]. The approach works on single slices but over time so analyzing the flow of a contrast agent. This time component adds the third dimension and shows to increase the performance of the classification in an important way. The variety of applications show the large spectrum of applications in multidimensional data.

2.4 Visual Features

Visual features remain important, particularly transferring visual features from other domains to the use within medical imaging. In many benchmarks it was shown that the more visual features are used the better the results are [9]. Still, when moving towards big data it will simply become impracticable to work with too large a variety of features and thus optimized and compact visual features will become necessary.

In *customised frequency pre–filtering in a local binary pattern–based classification of gastrointestinal images*, Wimmer et al. present the use of Local Binary Patterns (LBP) for the analysis of gastrointestinal images [22]. LBPs have been used in various scenarios to represent texture information. In the case of gastrointestinal images (in this case for Celiac disease and polyps) a frequency–based pre–filtering of the images led to optimized results that outperformed LBP and several of its derivations.

Garcia Seco de Herrera et al. describe in *bag of colors for biomedical document image classification* the use of SIFT and bag of visual color features for document image classification in images types [7]. Using a standard data set of the ImageCLEF benchmark the two features combined showed to increase the classification performance more than any of the participants in the benchmark that only used the supplied training data. This shows that color carries a very important part of the document image information.

2.5 Multimodal Retrieval

The last session of the day before the panel was on multimodal retrieval approaches, meaning in this case the combination of visual retrieval techniques and textual retrieval and not the combination of modalities such as PET and CT or MRI.

In *an SVD–bypass latent semantic analysis for image retrieval*, Stathopoulos et al. use latent semantic indexing on visual and textual information for image retrieval [20]. The results of mixing the modalities performs well on the given ImageCLEF database. Another aspect of the search was the scalability, using simple visual features that could potentially scale to millions of images for the retrieval.

Castellanos et al. describe in *multimedia retrieval in a medical image collection: results using modality classes* an approach for using modality classes for retrieval from the ImageCLEF 2011 medical task [3]. Expanding queries with an automatically extracted image modality class overall improves the results. These gains strongly depend on the type of query and it can not be generalized to all types of queries. Modality class information can also be used to improve the figure captions or their context, so the corresponding figures can be retrieved easier in the future.

3 The Panel and Discussions

The presentations throughout the day have demonstrated the large variety in applications, in image types (CT, x–ray, MRI, dual energy CT, ultrasound, gastrointestinal videos, journal figures, ...) and also in techniques and visual features used. Medical image retrieval has found its way as part of several medical applications and will often do its work as part of a larger system, whether it is combined with text search, for getting decision support or in getting access to interesting cases for preparing courses.

The panel was lead by Tanveer Syeda-Mahmood as an open session around the topic: *What is the CBIR role in Medical Decision Support?*. The panel started with an experience report of IBM on applications of medical image retrieval in working with clients on various integration projects. Several issues were raised including the difficulty to actually have clinically relevant tools and systems that help physicians in doing their work better and quicker. Pure system performance in benchmarks does not correlate with user acceptance and many discussions and tests with the physicians were necessary to conserve only those parts of the systems that really have an added value for the physician. This means that the techniques need to be integrated in the work flow, they need to be fast and in many cases they will be invisible.

The open panel invited all participants of the workshop to join the discussion and propose ideas. One issue that was discussed is the need for high quality *annotated medical imaging data* in large quantities. Having access to such data would be a big advantage for the community, but those who prepare the data would need to get at least part of the benefits, which is currently not always the case. This can lead to data sets not being shared. Quality of the annotation and confidence in diagnosis were also mentioned as physicians often are confident of their own opinion and do not necessarily trust other physicians unless they have a sort of proof, for example in the form of biopsies. Inter rater disagreement is important to measure, also to have a baseline for computer–based decision support. The level of detail for annotation can vary strongly and this needs to be defined well to create useful data sets and not have essential information missing. Community efforts are expected to be needed where several partners create annotations together. It was also mentioned that there will always be new needs for data sets as benchmarking has also the risk to lead to standardization and block new ideas.

The importance of *big data* was stressed as this could allow totally different approaches and maybe would lead with much simpler techniques to better results. Current computers have the possibility to deal with such extremely large amounts of data and much can still be discovered in this respect. Data protection of course needs to be respected and informed consent is in most cases necessary, even though data protection is different depending on the countries even inside Europe. Clear guidelines for the secondary use of medical imaging data in anonymized form need to be developed to ease research all while respecting privacy of the patients.

4 Conclusions

Medical image retrieval has remained a very dynamic research domain over the past ten years with many new directions and a strong evolution. Focus has come from theoretical models towards real applications and from small data sets to much larger data repositories. Whereas initial image retrieval applications focused on the image and human perception of visual similarity, modern applications are increasingly integrating the medical context into the retrieval process, such as mixing visual image data with clinical parameters, or several images for real case–based retrieval to help diagnosis and create links between a clinical case and reports in the medical literature. Many 3D and even 4D retrieval applications have started. More than in 2D retrieval, these applications will require the definitions of regions of interest to focus a more detailed analysis on the most important parts. Detecting small regions of potential abnormalities and then using these regions to find images or cases with similar lesions will remain an important direction for the coming years to improve current tools for medical decision support. Models for organs need to be built and links between images from the scientific literature in JPEG format and in the DICOM format in the patient record need to be available as well, to make sure that the various sources of information are in the end well connected.

Image retrieval based on visual and textual data might not be a visible part of all applications, but increasingly it is integrated into many tools, even though often in an invisible form. This indicates that some of the search and retrieval research should be conducted as part of specific applications or specific domains. Much can still be improved in terms of techniques and scalable approaches to deal with big data but it can be foreseen that several of the tools will be integrated into a variety of applications and systems, with potential commercial success.

Acknowledgments. This work was supported by the EU in the FP7 through the VISCERAL (318068), PROMISE (258191) and Khresmoi (257528) projects.

References

1. Riding the wave: How europe can gain from the rising tide of scientific data. Submission to the European Comission (October 2010), http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
2. Akgül, C., Rubin, D., Napel, S., Beaulieu, C., Greenspan, H., Acar, B.: Content–based image retrieval in radiology: Current status and future directions. Journal of Digital Imaging 24(2), 208–222 (2011)
3. Castellanos, A., Benavent, X., García-Serrano, A., Cigarrán, J.: Multimedia Retrieval in a Medical Image Collection: Results Using Modality Classes. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 133–144. Springer, Heidelberg (2013)

4. Catalano, C.E., Robbiano, F., Parascandolo, P., Cesario, L., Vosilla, L., Barbieri, F., Spagnuolo, M., Viano, G., Cimmino, M.A.: Exploiting 3D Part-Based Analysis, Description and Indexing to Support Medical Applications. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 21–32. Springer, Heidelberg (2013)
5. El-Kwae, E., Xu, H., Kabuka, M.R.: Content–based retrieval in picture archiving and communication systems. JDI 13(2), 70–81 (2000)
6. Foncubierta–Rodríguez, A., Vargas, A., Platon, A., Poletti, P.–A., Müller, H., Depeursinge, A.: Retrieval of 4D Dual Energy CT for Pulmonary Embolism Diagnosis. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 45–55. Springer, Heidelberg (2013)
7. de Herrera, A.G.S., Markonis, D., Müller, H.: Bag–of–Colors for Biomedical Document Image Classification. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 110–121. Springer, Heidelberg (2013)
8. Hersh, W., Müller, H., Kalpathy-Cramer, J., Kim, E., Zhou, X.: The consolidated ImageCLEFmed medical image retrieval task test collection. Journal of Digital Imaging 22(6), 648–655 (2009)
9. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum) (September 2011)
10. Kato, T.: Database architecture for content–based image retrieval. In: Jamberdino, A.A., Niblack, W. (eds.) Image Storage and Retrieval Systems. SPIE Proc., San Jose, California, vol. 1662, pp. 112–123 (February 1992)
11. Langs, G., Hanbury, A., Menze, B., Müller, H.: VISCERAL: Towards Large Data in Medical Imaging — Challenges and Directions. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 92–98. Springer, Heidelberg (2013)
12. Lowe, H.J., Antipov, I., Hersh, W., Smith, C.A.: Towards knowledge–based retrieval of medical images. The role of semantic indexing, image content representation and knowledge–based retrieval. In: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Nashville, TN, USA, pp. 882–886 (October 1998)
13. Müller, H., Clough, P., Hersh, W., Deselaers, T., Lehmann, T., Geissbuhler, A.: Evaluation axes for medical image retrieval systems: the ImageCLEF experience. In: MULTIMEDIA 2005: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 1014–1022. ACM Press, New York (2005)
14. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine–clinical benefits and future directions. International Journal of Medical Informatics 73(1), 1–23 (2004)
15. Qian, Y., Wang, L., Wang, C., Gao, X.: The Synergy of 3D SIFT and Sparse Codes for Classification of Viewpoints from Echocardiogram Videos. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 68–79. Springer, Heidelberg (2013)
16. Quatrehomme, A., Millet, I., Hoa, D., Subsol, G., Puech, W.: Assessing the Classification of Liver Focal Lesions by Using Multi-phase Computer Tomography Scans. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 80–91. Springer, Heidelberg (2013)

17. Quellec, G., Lamard, M., Droueche, Z., Cochener, B., Roux, C., Cazuguel, G.: A Polynomial Model of Surgical Gestures for Real-Time Retrieval of Surgery Videos. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 10–20. Springer, Heidelberg (2013)
18. Simonyan, K., Modat, M., Ourselin, S., Cash, D., Criminisi, A., Zisserman, A.: Immediate ROI Search for 3-D Medical Images. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 56–67. Springer, Heidelberg (2013)
19. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
20. Stathopoulos, S., Kalamboukis, T.: An SVD–Bypass Latent Semantic Analysis for Image Retrieval. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 122–132. Springer, Heidelberg (2013)
21. Tagare, H.D., Jaffe, C., Duncan, J.: Medical image databases: A content–based retrieval approach. Journal of the American Medical Informatics Association 4(3), 184–198 (1997)
22. Hegenbart, S., Maimone, S., Uhl, A., Vécsei, A., Wimmer, G.: Customised Frequency Pre-Filtering in a Local Binary Pattern-Based Classification of Gastrointestinal Images. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 99–109. Springer, Heidelberg (2013)
23. Yang, S., Shapiro, L., Cunningham, M., Speltz, M., Birgfeld, C., Atmosukarto, I., Lee, S.-I.: Skull Retrieval for Craniosynostosis Using Sparse Logistic Regression Models. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 33–44. Springer, Heidelberg (2013)

A Polynomial Model of Surgical Gestures for Real-Time Retrieval of Surgery Videos

Gwénolé Quellec[1], Mathieu Lamard[1,2], Zakarya Droueche[1,3],
Béatrice Cochener[1,2,4], Christian Roux[1,3], and Guy Cazuguel[1,3]

[1] Inserm, UMR 1101, Brest, F-29200 France
gwenole.quellec@inserm.fr
[2] Univ Bretagne Occidentale, Brest, F-29200 France
mathieu.lamard@univ-brest.fr
[3] INSTITUT. TELECOM, TELECOM Bretagne, UEB,
Dpt. ITI, Brest, F-29200 France
{mohammed.droueche,christian.roux,guy.cazuguel}@telecom-bretagne.eu
[4] CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France
beatrice.cochener@ophtalmologie-chu29.fr

Abstract. This paper introduces a novel retrieval framework for surgery videos. Given a query video, the goal is to retrieve videos in which similar surgical gestures appear. In this framework, the motion content of short video subsequences is modeled, in real-time, using spatiotemporal polynomials. The retrieval engine needs to be trained: key spatiotemporal polynomials, characterizing semantically-relevant surgical gestures, are identified through multiple-instance learning. Then, videos are compared in a high-level space spanned by these key spatiotemporal polynomials. The framework was applied to a dataset of 900 manually-delimited clips from 100 cataract surgery videos. High classification performance ($A_z = 0.816 \pm 0.118$) and retrieval performance ($MAP = 0.358$) were observed.

Keywords: surgery video retrieval, motion analysis, spatiotemporal polynomials, multiple-instance learning.

1 Introduction

Automated analysis of videos, in the context of video-monitored surgery, is an increasingly active research field. Several methods have been proposed to finely analyze regions of interest (through image mosaicing) [1], to detect surgical tools (for augmented reality purposes) [2], to categorize surgical tasks [3], or to identify key surgical events [4]. In line with these works, we are developing a retrieval system for surgery videos. This system could be used to help surgeons browse large video archives, for retrospective studies or for training purposes. It could also be used to generate warnings and recommendations, in real-time, during a video-monitored surgery.

The purpose of video retrieval systems [5] is to find, within digital archives, videos that resemble a query video. Initially popularized in broadcasting [6] and

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 10–20, 2013.

video surveillance [7,8], it recently started developing in other applications. For instance, its use for medical training has been considered [9].

Typical video retrieval systems accept a video file as input and display similar video files on output [10]. To compare two videos, several feature vectors are extracted from each video. These feature vectors typically represent the shape, the texture, the color or, more importantly, the motion content of each video at different time instants. Motion feature extraction usually involves motion segmentation [11,12] or salient point characterization [13,14]. Segmenting the motion content of a surgery video is challenging, partly because many moving objects have fuzzy borders and are deformable. Detecting salient point, on the other hand, is possible, but useful information does not necessarily lie where salient points are detected. A different solution is proposed: the motion content of short video subsequences is characterized globally, using a deformable motion model; a polynomial model was adapted. A few authors proposed the use of spatial polynomials [15,16] or spatiotemporal polynomials [17,18] for motion analysis. However, the order of the spatiotemporal polynomials was limited to 2 [17] or 3 [18]; a generalization to arbitrary spatiotemporal polynomial orders is proposed in this paper.

Once feature vectors have been extracted, videos are generally compared using a global similarity metric. Practical challenges arise when comparing two videos globally. In particular, there can be large variations of speed within semantically-similar videos. Dynamic Time Warping (DTW) and its derivatives [19,10] were specifically designed to overcome this challenge. Other solutions rely on Hidden-Markov Models (HMM) [20,21]. However, in a surgery video, not all surgical gestures are semantically relevant. In eye surgery, for instance, surgeons have to moisturize the eye at regular intervals, regardless of the current surgical task. Moreover, the motion content is not only caused by surgical gestures: it is partly caused by camera motion, patient motion, etc. In order to define a semantically-relevant similarity metric between videos, we need to consider only those video subsequences in which a semantically-relevant gesture has been detected. Selecting key surgical gestures is not possible in the solutions above, so a different solution was proposed, as described hereafter. This solution relies on the boosted Diverse Density algorithm [22].

2 Overview of the Method

In this framework, the motion content of short video subsequences is modeled, in real-time, using spatiotemporal polynomials (section 3). Then, key spatiotemporal polynomials, characterizing semantically-relevant surgical gestures, are identified through multiple-instance learning (section 4). Finally, a distance between videos is defined. In order to retrieve semantically-similar videos, this distance is computed in a space spanned by these key spatiotemporal polynomials (section 5).

3 Motion Characterization in Video Subsequences

Let $V^{(i)} = \left\{V_1^{(i)}, V_2^{(i)}, ..., V_{n_i}^{(i)}\right\}$ denote a video sequence of n_i frames. Videos are divided into possibly overlapping subsequences of n frames; one video subsequence is analyzed every m frames. Let $V_{[jm+1;jm+n]}^{(i)} = \left\{V_{jm+1}^{(i)}, V_{jm+2}^{(i)}, ..., V_{jm+n}^{(i)}\right\}$ denote the j^{th} video subsequence in $V^{(i)}$. Parameter m is chosen to tradeoff retrieval precision and computation times. Note that subsequences overlap if and only if $m < n$.

In each video subsequence, the motion content is extracted as described in section 3.1 and characterized as described in section 3.2.

3.1 Motion Extraction

The motion content of a video subsequence, say $V_{[jm+1;jm+n]}^{(i)}$, is extracted from the optical flow between pairs of consecutive frames. Let $V_k^{(i)}$ and $V_{k+1}^{(i)}$ denote two consecutive frames in $V_{[jm+1;jm+n]}^{(i)}$.

To compute the optical flow between $V_k^{(i)}$ and $V_{k+1}^{(i)}$, D strong corners are first detected in $V_k^{(i)}$. Strong corners are a particular type of salient points that are supposedly easy to track. They are defined as the pixels $(x_{k\pi}, y_{k\pi})$ of frame $V_k^{(i)}$ that maximize the smallest eigenvalue of matrix $M_{k\pi}$ below:

$$M_{k\pi} = \begin{pmatrix} \sum_{(x,y)\in\mathcal{V}_{k\pi}} \left(\frac{\partial V_k^{(i)}}{\partial x}(x,y)\right)^2 & \sum_{(x,y)\in\mathcal{V}_{k\pi}} \frac{\partial V_k^{(i)}}{\partial x}(x,y)\cdot\frac{\partial V_k^{(i)}}{\partial y}(x,y) \\ \sum_{(x,y)\in\mathcal{V}_{k\pi}} \frac{\partial V_k^{(i)}}{\partial y}(x,y)\cdot\frac{\partial V_k^{(i)}}{\partial x}(x,y) & \sum_{(x,y)\in\mathcal{V}_{k\pi}} \left(\frac{\partial V_k^{(i)}}{\partial y}(x,y)\right)^2 \end{pmatrix} \tag{1}$$

where $\mathcal{V}_{k\pi}$ is a neighborhood of pixel $(x_{k\pi}, y_{k\pi})$. The optical flow between $V_k^{(i)}$ and $V_{k+1}^{(i)}$ is then computed, at each strong corner, using the Lucas-Kanade iterative method [23]. On output, a motion field $\mathcal{F}_{ij}$ is obtained. Each element $f_d \in \mathcal{F}_{ij}$ maps a spatiotemporal coordinate ($x_d = x_{k\pi}$, $y_d = y_{k\pi}$, $t_d = k - jm$) to a displacement (u_d, v_d), $d = 1..D$.

Just because they are salient, the detected strong corners are not expected to be clinically relevant. What is clinically relevant is the overall motion field within the video subsequence. The strong corners, and the displacement measured at their locations, simply are convenient support vectors of that motion field $\mathcal{F}_{ij}$.

3.2 Motion Characterization Using Spatiotemporal Polynomials

In order to characterize the motion field within subsequence $V_{[jm+1;jm+n]}^{(i)}$, the motion vectors in $\mathcal{F}_{ij}$ are approximated by two spatiotemporal polynomials. The first polynomial maps the spatiotemporal coordinate (x, y, t) to the horizontal displacement u. The second maps the spatiotemporal coordinate to the vertical displacement v.

Let p denote the maximal polynomial order. Given a basis of canonical polynomials, noted $\mathcal{C}_p$, we search for the polynomial coefficients that minimize the sum of the squared errors between the true motion field $\mathcal{F}_{ij}$ and the motion field approximated by a polynomial model of order p. These optimal polynomial coefficients are the motion signature of subsequence $V^{(i)}_{[jm+1;jm+n]}$.

Polynomial bases have the following structure: $\mathcal{C}_1 = \{1, x, y, t\}$, $\mathcal{C}_2 = \{1, x, y, t, xy, xt, yt, x^2, y^2, t^2\}$, etc. Let $L_p = |\mathcal{C}_p|$ denote the number of canonical polynomials: $L_1 = 4$, $L_2 = 10$, $L_3 = 20$, $L_4 = 35$, $L_5 = 56$, etc.

Let C_{pd} denote the vector formed by the canonical vectors in $\mathcal{C}_p$, evaluated at coordinate (x_d, y_d, t_d), $d = 1..D$; for instance, $C_{1d} = (1, x_d, y_d, t_d)$. The approximated motion vector can be expressed as a matrix product $C_{pd}P_{pij}$, where $P_{pij} \in \mathcal{M}_{L_p,2}$ contains the optimal polynomial coefficients (see Fig. 1). Matrix P_d is defined as follows:

$$P_{pij} = \underset{X}{\operatorname{argmin}} \sum_{d=1}^{D} \left\|(u_d, v_d) - C_{pd}X\right\|^2 \qquad (2)$$

The optimal solution is found when the derivative of the sum, with respect to matrix X, equals 0. This solution can be rewritten as follows:

$$\begin{cases} A_p P_{pij} = E_p \\ A_p = \sum_{d=1}^{D} C_{pd}^T C_{pd} \\ E_p = \sum_{d=1}^{D} C_{pd}^T (u_d, v_d) \end{cases} \qquad (3)$$

with $A_p \in \mathcal{M}_{L_p,L_p}$ and $E_p \in \mathcal{M}_{L_p,2}$. Matrix P_{pij} is obtained by solving two systems of linear equations of order L_p: one for the horizontal displacements and one for the vertical displacements. In the first system, the right-hand side of the equation is the first column of E_p. In the second system, the right-hand side is the second column of E_p. These systems are solved using the LU decomposition of A_p. The solution of the first (respectively the second) system is stored in the first (respectively the second) column of P_{pij}.

3.3 Complexity Analysis

The most complex step in the computation of P_{pij}, given the motion field $\mathcal{F}_{ij}$, is the computation of matrix A_p. Since A_p is symmetric, its computation requires $O(\frac{1}{2}DL_p^3)$ operations. In comparison, the complexity of its LU decomposition is in $O(\frac{1}{2}L_p^3)$. So, the complexity of the entire process increases linearly with D, the number of detected strong corners (section 3.1).

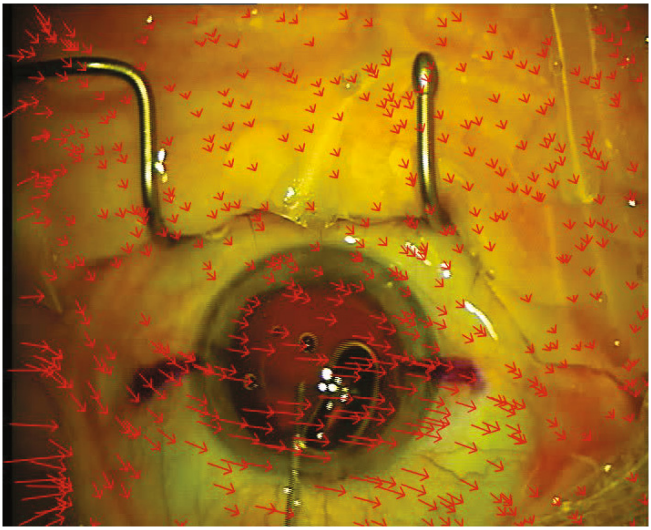


Fig. 1. Motion field approximated in one frame, at randomly-selected coordinates, using the polynomial model (polynomial order p=6, subsequence length: n=25 frames)

4 Key Surgical Gesture Selection

Once each video subsequence has been characterized, multiple-instance learners are used to identify key spatiotemporal polynomials, i.e. spatiotemporal polynomials characterizing key surgical gestures.

4.1 Multiple-Instance Learning

Multiple-instance learners are supervised learners that receive a set of *bags of instances*. A binary label (relevant or irrelevant) is assigned to each bag [24]. A bag is labeled irrelevant if all the instances in it are irrelevant. On the other hand, a bag is labeled relevant if it contains at least one relevant instance (or one key instance). From a collection of labeled bags, multiple-instance learners are trained to detect relevant instances.

In this paper, each video subsequence $V^{(i)}_{[jm+1;jm+n]}$ (or more exactly its signature P_{pij}) is regarded as an instance. Each video $V^{(i)}$ is regarded as a bag of instances. If there are more than two class labels, Multiple-Instance Learning (MIL) is applied separately to each class label γ, $\gamma = 1..\Gamma$: a video is relevant if and only if its original class label is γ. Irrelevant and relevant videos are noted $V^{(i,-)}$ and $V^{(i,+)}$, respectively, and the signatures of their subsequences are noted P^{-}_{pij} and P^{+}_{pij}, respectively.

The most popular MIL frameworks are Andrews' SVM [25], Diverse Density [24], and their derivatives. In Andrews' SVM, a support-vector machine processes the instance labels as unobserved integer variables, subjected to constraints defined by the bag labels. The goal is to maximize the soft-margin over hidden label variables and a discriminant function. Diverse Density measures the intersection of the relevant bags minus the union of the irrelevant bags. The location of relevant instance in feature space, and also the best weighting of the features, is found by maximizing Diverse Density. Diverse Density was chosen for its simplicity and its generality: in particular, it will allow spatiotemporal polynomial basis adaptation in future works (section 8).

4.2 Diverse Density

Diverse Density, in its simplest form, is defined as follows:

$$\begin{cases} \widehat{P} = \underset{P}{\operatorname{argmax}} \prod_i Pr(+|V^{(i,+)}, P) \prod_i Pr(-|V^{(i,-)}, P) \\ Pr(+|V^{(i,+)}, P) = 1 - \prod_j \left[1 - \exp\left(-\left\|P^+_{pij} - P\right\|^2\right)\right] \\ Pr(-|V^{(i,-)}, P) = \prod_j \left[1 - \exp\left(-\left\|P^-_{pij} - P\right\|^2\right)\right] \end{cases} \quad (4)$$

$\widehat{P}$, the key spatiotemporal polynomial, is found by several gradient ascents controlled by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [26]. Each ascent is initialized by an instance randomly selected within the relevant bags and the best solution is retained. Features can be weighted simultaneously: the Euclidean distance simply needs to be replaced by a weighted Euclidean distance [24]:

$$\begin{cases} (\widehat{P}, \widehat{s}) = \underset{P,s}{\operatorname{argmax}} \prod_i Pr(+|V^{(i,+)}, P, s) \prod_i Pr(-|V^{(i,-)}, P, s) \\ \left\|P_{pij} - P\right\|^2 = \sum_{k=1}^{L_p} \sum_{l=1}^{2} s_{kl}^2 (P_{pijkl} - P_{kl})^2 \end{cases} \quad (5)$$

The weights of the Euclidean distance are also found using the gradient ascent.

4.3 Boosted Diverse Density

Note that there may be several key surgical gestures in each class. Typically, a surgical task may be composed of several sub-tasks. Besides, two surgeons may perform the same surgical task using different techniques. In order to select several key surgical gestures when need be, a boosted version of Diverse Density was used [22]. Let $\widehat{P}_{\gamma b}$, $b = 1..B_\gamma$, be the key spatiotemporal polynomials found by boosted Diverse Density for class γ, $\gamma = 1..\Gamma$, and let $\widehat{s}_{\gamma b}$ be the associated feature weights.

5 Distance Metric

Let U and V be two videos characterized as described in section 3. We would like to define a semantically-relevant distance $D(U, V)$ between these videos. To be semantically-relevant, this distance should ignore spatiotemporal polynomials that are too far away from the key spatiotemporal polynomials found in section 4.

A new space is defined: in that space, component $U_{\gamma b}$ (respectively $V_{\gamma b}$) conveys the probability that key spatiotemporal polynomial $\widehat{P}_{\gamma b}$ is present in video U (respectively video V):

$$U_{\gamma b} = Pr(+|U, \widehat{P}_{\gamma b}, \widehat{s}_{\gamma b}) \tag{6}$$

$D(U, V)$ simply is the Euclidean distance in that space:

$$D(U, V) = \sqrt{\sum_{\gamma=1}^{\Gamma} \sum_{b=1}^{B_\gamma} (U_{\gamma b} - V_{\gamma b})^2} \tag{7}$$

This distance metric is now used for video retrieval in a dataset of cataract surgery videos.

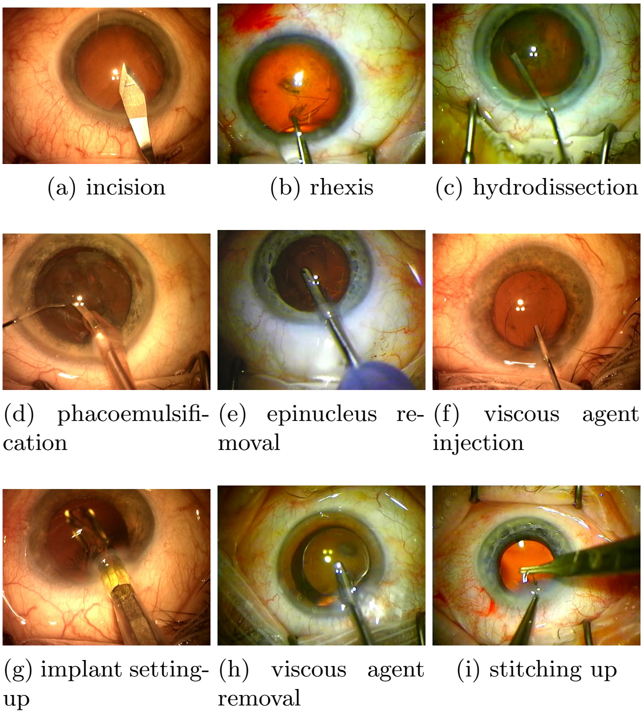
6 Cataract Surgery Dataset

A dataset of 100 videos from 100 consecutive cataract surgeries was collected at Brest University Hospital (Brest, France) between February and July 2011. Surgeries were performed by ten different surgeons. Some videos were recorded with a CCD-IRIS device (Sony, Tokyo, Japan), the others were recorded with a MediCap USB200 video recorder (MediCapture, Philadelphia, USA). They were stored in MPEG2 format, with the highest quality settings, or in DV format. Image definition is 720 x 576 pixels. The frame frequency is 25 frame per seconds.

In each video, a temporal segmentation was provided by cataract experts for each surgical task. The following surgical tasks were temporally segmented in videos: incision, rhexis, hydrodissection, phacoemulsification, epinucleus removal, viscous agent injection, implant setting-up, viscous agent removal and stitching up (see Fig. 2). As a result, nine manually-delimited clips were obtained per surgery. Overall, 900 clips were obtained: the first 450 clips (obtained from the first 50 surgeries) were assigned to the training set, the last 450 clips were assigned to the test set. Clips have an average duration of 94 seconds (standard deviation: 77 seconds, min: 9 seconds, max: 312 seconds).

7 Results

The proposed system has four parameters (section 3): the number of frames per subsequence (n), the delay between the beginning of two consecutive subsequences (m), the number of selected strong corners per frame (D) and the maximal polynomial order (p). Two parameters were chosen empirically to allow real-time fitting of the polynomial models: m=5 frames and D=400 strong



(a) incision (b) rhexis (c) hydrodissection

(d) phacoemulsification (e) epinucleus removal (f) viscous agent injection

(g) implant setting-up (h) viscous agent removal (i) stitching up

Fig. 2. High-level surgical tasks

corners per frame. The other two (n and p) were chosen by two-fold cross-validation in the training set. Each tested (n, p) pair was graded by the average, over all surgical tasks $\gamma = 1..9$ and both folds, of the area under the Receiver Operating Characteristic (ROC) of $\sum_{b=1}^{B_\gamma} Pr(+|V, \widehat{P}_{\gamma b}, \widehat{s}_{\gamma b})$ (see equation 5). The optimal pair was $(n = 3, p = 25)$. For each surgical task $\gamma = 1..9$, the ROC curve of $\sum_{b=1}^{B_\gamma} Pr(+|V, \widehat{P}_{\gamma b}, \widehat{s}_{\gamma b})$, in the test set, is reported in Fig. 3.

To evaluate the proposed distance measure (section 5), the Mean Average Precision (MAP) was measured. To measure the MAP, test videos was used one by one as the query. The nearest neighbors of each query were searched in the training set with respect to the proposed distance measure. A retrieved video was regarded as relevant if it belongs to the same task as the query video. The average precision at 5, 10, 20 and 50 was 0.561, 0.462, 0.401 and 0.179, respectively. A MAP of 0.358 was obtained. In comparison, a MAP of 0.287 was obtained using Dynamic-Time Warping [19] with the same subsequence signatures.

The system could process 29 frames per second on one core of an Intel Xeon processor running at 2.53 GHz (real-time: 25 frames per second).

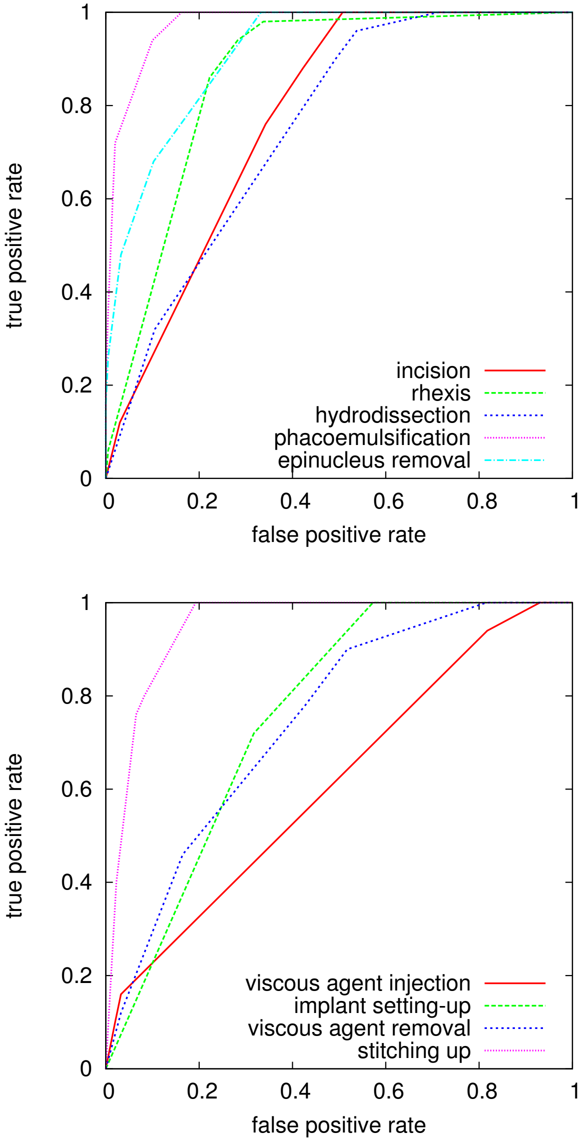


Fig. 3. Performance of key surgical gesture selection

8 Discussion

A video retrieval framework was presented in this paper. The motion content of short video subsequences was modeled using spatiotemporal polynomials. Then, key spatiotemporal polynomials were identified through multiple-instance learning. Finally, a distance between videos was defined in a high-level space spanned by these key spatiotemporal polynomials.

The entire process is very fast. Firstly, the spatiotemporal polynomials are computed on the fly, in real-time. Secondly, finding the location of a video in the high-level space above can also be done progressively, in real-time, as the subsequences are characterized. Thirdly, the proposed distance metric simply is a Euclidean distance in that high-level space, so finding the nearest neighbors of a video is easily scalable, using efficient search structures.

In this paper, the basis of spatiotemporal polynomials consisted of canonical polynomials. In future works, the basis will be adapted to push performance further: each basis will be a linear combination of canonical polynomials. The coefficients of these linear combinations could be tuned to maximize Diverse Density through gradient ascents: this is what motivated the selection of Diverse Density as a multiple-instance learner.

The proposed framework seems particularly suited to surgery videos. This was checked in a cataract surgery dataset. Experiments on other types of surgery would be needed to confirm. Moreover, this framework could be beneficial to other problems where motion cannot be easily segmented and where relevant visual information does not necessarily lie in the neighborhood of salient points.

References

1. Seshamani, S., Lau, W., Hager, G.: Real-Time Endoscopic Mosaicking. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 355–363. Springer, Heidelberg (2006)
2. Cano, A.M., Gayá, F., Lamata, P., Sánchez-González, P., Gómez, E.J.: Laparoscopic Tool Tracking Method for Augmented Reality Surgical Applications. In: Bello, F., Edwards, E. (eds.) ISBMS 2008. LNCS, vol. 5104, pp. 191–196. Springer, Heidelberg (2008)
3. Cao, Y., Liu, D., Tavanapong, W., Wong, J., Oh, J., de Groen, P.: Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos. IEEE Trans. Biomed. Eng. 54(7), 1268–1279 (2007)
4. Giannarou, S., Yang, G.-Z.: Content-Based Surgical Workflow Representation Using Probabilistic Motion Modeling. In: Liao, H., Edwards, P.J., Pan, X., Fan, Y., Yang, G.-Z. (eds.) MIAR 2010. LNCS, vol. 6326, pp. 314–323. Springer, Heidelberg (2010)
5. Patel, B.V., Meshram, B.B.: Content based video retrieval systems. Int. J. UbiComp 3(2), 13–30 (2012)
6. Naturel, X., Gros, P.: Detecting repeats for video structuring. Multimedia Tools and Applications 38(2), 233–252 (2008)
7. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: 8th ACM Int. Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York (2006)
8. Hu, W., Xie, D., Fu, Z., Zeng, W., Maybank, S.: Semantic-based surveillance video retrieval. IEEE Trans. Image. Process. 16(4), 1168–1181 (2007)
9. André, B., Vercauteren, T., Buchner, A.M., Shahid, M.W., Wallace, M.B., Ayache, N.: An Image Retrieval Approach to Setup Difficulty Levels in Training Systems for Endomicroscopy Diagnosis. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 480–487. Springer, Heidelberg (2010)

10. Xu, D., Chang, S.F.: Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans. Pattern. Anal. Mach. Intell. 30(11), 1985–1997 (2008)
11. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: Segmenting, modeling, and matching video clips containing multiple moving objects. IEEE Trans. Pattern. Anal. Mach. Intell. 29(3), 477–491 (2007)
12. Yamasaki, T., Aizawa, K.: Motion segmentation and retrieval for 3d video based on modified shape distribution. EURASIP J. Appl. Signal. Process 2007(1), 059535 (2007)
13. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. 64(2-3), 107–123 (2005)
14. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: ACM Int. Conf. on Image and Video Retrieval, Amsterdam, The Netherlands, pp. 494–501 (2007)
15. Jeannin, S.: On the combination of a polynomial motion estimation with a hierarchical segmentation based video coding scheme. In: Int. Conf. on Image Processing, Lausanne, Switzerland, pp. 489–492 (1996)
16. Kihl, O., Tremblais, B., Augereau, B., Khoudeir, M.: Human activities discrimination with motion approximation in polynomial bases. In: Int. Conf. on Image Processing, Hong Kong, China, pp. 2469–2472 (2010)
17. Hu, X., Ahuja, N.: Long image sequence motion analysis using polynomial motion models. In: IAPR Workshop on Machine Vision Applications, Tokyo, Japan, pp. 109–114 (1992)
18. Jakubiak, J., Nomm, S., Vain, J., Miyawaki, F.: Polynomial based approach in analysis and detection of surgeon's motions. In: Int. Conf. on Control, Automation, Robotics and Vision, Hanoi, Vietnam, pp. 611–616 (2008)
19. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. 26(1), 43–49 (1978)
20. Lee, D.S.: Meeting video retrieval using dynamic HMM model similarity. In: IEEE Int. Conf. on Multimedia and Expo., Amsterdam, The Netherlands (July 2005)
21. Lili, N.A.: Hidden markov model for content-based video retrieval. In: Asia Int. Conf. on Modelling and Simulation, Bandung, Indonesia, pp. 353–358 (2009)
22. Foulds, J.R., Frank, E.: Speeding up and boosting diverse density learning. In: Conf. on Discovery Science, Lyon, France, pp. 102–116 (2010)
23. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: DARPA Imaging Understanding Workshop, Washington, DC, USA, pp. 121–130 (1981)
24. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Conf. Advances in Neural Information Processing Systems, pp. 570–576. Denver, Co., USA (1998)
25. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, Whistler, Canada, vol. 15, pp. 561–568 (2003)
26. Broyden, C.G.: The convergence of a class of double-rank minimization algorithms. J. Inst. Math. Appl. 6, 76–90 (1970)

Exploiting 3D Part-Based Analysis, Description and Indexing to Support Medical Applications

Chiara Eva Catalano[1], Francesco Robbiano[1], Patrizia Parascandolo[2],
Lorenzo Cesario[2], Loris Vosilla[2], Francesca Barbieri[3], Michela Spagnuolo[1],
Gianni Viano[2], and Marco Amedeo Cimmino[3]

[1] CNR IMATI-Genova, Via De Marini 6, Genova
{chiara.catalano,francesco.robbiano,
michela.spagnuolo}@ge.imati.cnr.it
[2] Softeco Sismat S.r.l., Via De Marini 1, Genova
{patrizia.parascandolo,lorenzo.cesario,loris.vosilla,
gianni.viano}@softeco.it
[3] DIMI, Dipartimento di Medicina Interna, Clinica Reumatologica,
Università degli Studi di Genova
{francesca.barbieri,cimmino}@unige.it

Abstract. Multi-modality is crucial to handle knowledge, medical data and patient-specific information in an integrated fashion: in the course of their work, indeed, clinicians analyze a large amount of information about interrelated layers such as anatomy, kinematics, dynamics, mechanics and physiology. Much of the information related to these levels is intrinsically 3D and we believe that the adoption of 3D part-based annotation and content-based indexing will open up new ways to integrate and interact with medical information. In this paper, we will focus the attention on content-based analysis of 3D medical data and discuss related issues and trends, based on two software tools: the ShapeAnnotator and RheumaSCORE. In the illustrative scenario of the Rheumatoid Arthritis we will provide hints for even more informative Computer Aided Diagnosis systems for clinical support.

Keywords: 3D shape segmentation, 3D part-based annotation, 3D content-based description, Computer Aided Diagnosis, Rheumatoid Arthritis.

1 Introduction

With the rapid innovation in computing and electronic imaging technology, there has been increasing interest in developing Computer Aided Detection/Diagnosis (CAD) systems to improve the medical service [1,2]. CAD is emerging as an advanced interdisciplinary technology, which combines fundamental elements of different areas such as digital image processing, image analysis, pattern recognition, medical information processing and knowledge management.

CAD systems are likely to become the means of processing and interaction of the huge amount of available digital data and to incorporate new methods for comparative

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 21–32, 2013.

analysis and study of clinical cases. Methods of information retrieval will play a fundamental role in offering techniques to correlate and analyse such data based on qualitative and quantitative information extracted automatically. Content-based medical retrieval systems have the potential to improve the performance of clinicians [3] speeding up the diagnosis process and improving accuracy and treatment of complex diseases.

While the importance of content-based techniques is not new, this paper focuses the attention on issues related to content-based analysis of *3D* data in the medical domain. The success of CAD-supported analysis processes depends indeed on the capabilities of automated solutions to simulate and improve what physicians and radiologists do when they inspect digital data. We believe that the key challenges are:

- software applications should be able to *identify and measure* clinical parameters based on the *geometric/morphological characterization* of the shape of organs, anatomical elements or their parts;
- formal methods to assess the similarity among shapes should support the retrieval of *similar* clinical cases in order to speed up the diagnosis process and to support comparative analysis among known cases;
- gathering information about specific patients should ease the evaluation of their *follow-up* in order to highlight temporal trends of pathology markers, possibly depending on current therapy;
- performing *statistical analysis* over a significantly large population of patients would trigger the possible detection of new correlation patterns and speed up the screening of large populations for abnormal cases.

After the non-trivial segmentation and reconstruction steps, the usage of 3D models in the medical domain has mainly concerned the localization and visualization of parts. However, the highly informative content carried by 3D models can be exploited more heavily: 3D-based shape analysis, indexing, and part-based annotation can provide the novel means to integrate medical information in a truly multi-modal framework.

The contribution of this paper is the discussion about this integration, showing how 3D part-based characterization and annotation could enhance an existing CAD system. To substantiate our arguments, we will describe two existing tools, i.e. the ShapeAnnotator and RheumaSCORE, and demonstrate their complementarity and fruitful integration in the framework of Computer Assisted Diagnosis.

On the one side, the *ShapeAnnotator* is a 3D object annotation system that provides the framework for 3D part-based annotation, relying upon a multi-segmentation of 3D shapes and concepts formalized in an ontology. Along with fine-grained shape characterizations, annotated parts can be used to index relevant parts in 3D reconstructed models. On the other side, the *RheumaSCORE* software implements some prototypical tools for the automatic characterization of 3D parts, especially relying on the detection of bone erosion, and has been considered very useful to help radiologists or physicians during diagnostic processes and follow-up of rheumatic patients.

Using the diagnosis of Rheumatoid Arthritis (RA) as illustrative scenario, we will discuss how more informative content-based systems may be designed for clinical support.

The paper is organized as follows. In Section 2, the 3D multi-segmentation and annotation framework will be presented; some hints will be given on the 3D characterizations that can be automatically attached to parts in order to support their description and to be used for indexing purposes. In the medical scenario, the usage of these methodologies is still at its infancy, while we believe that coupling 3D analysis techniques with CAD systems has a high potential to innovate the field. In Section 3, the RA scenario will be described in the framework of the RheumaSCORE system and some clinical trials will be discussed. In Section 4, we will introduce an integration perspective, discussing on how these systems could be combined and improved, and briefly sketching future research directions. Section 5 concludes the paper with a short wrap-up.

2 3D Multi-segmentation, Annotation and Characterization

A variety of techniques have been developed in image processing for content analysis and segmentation: these are particularly useful in the medical domain where much of the digital content is stored as 3D images, including data from Computer Tomography (CT), Magnetic Resonance Imaging (MRI), MicroCT and other devices. The available techniques are generally based on the computation of low-level features (e.g. texture, color, edges) that are used to identify and isolate, or segment, relevant parts in the 2D or 3D image. Segmentation is primarily meant as the process to detect specific shapes in a 2D or 3D image, while content-based annotation methods allow to create correspondences between complex objects, or parts, and conceptual tags: once the content is analyzed and its relevant constituents annotated, they can easily match textual searches.

Getting accurate 3D reconstructions from raw segmented images is still very labor-intensive, but a growing number of problems in medical analysis require the manipulation of the full spatial geometry. In other words, 3D models represent the geometry of all the interesting parts (e.g. organs, bones) which can be further analyzed and possibly decomposed by dedicated computational tools [4]: in fact, 3D models, or their parts, may be characterized by morphological attributes, abstract properties (e.g. signatures for indexing and retrieval), or any other useful parameter.

With these premises, the ingredients needed to fully integrate 3D geometry into the medical processing pipeline are: 3D shape analysis tools, a rich set of tools to characterize the models and/or their parts, and a methodology to associate part-based annotations with the geometric representation. We emphasize that in the medical practice annotations need to be attached not only to the whole 3D reconstruction, but most frequently to *parts* of interest, for instance a bone or even a specific portion of it [5].

The *ShapeAnnotator* [6] was a first attempt to integrate segmentations and annotations for generic 3D models. It relies on the concurrent use of a variety of shape analysis and segmentation tools to offer a rich set of operators to detect 3D regions of interest on the 3D models. Given the complexity of shapes in the medical domain, it is widely recognized that no single algorithm can be used to provide a segmentation that yields interesting and exhaustive results for all feature types, even if the context is well-defined [7]. Hence, the ShapeAnnotator approached the problem of feature extraction via the concept of *multi-segmentation* of a 3D surface represented by a triangle mesh. The idea is to use in parallel a set of different segmentation procedures and to select and compose just the meaningful results from each of them: interesting features can be interactively selected from the resulting multi-segmented mesh.

The ShapeAnnotator allows loading an ontology to enrich the segmented model with concepts. To annotate the features of the 3D model, the user can choose the appropriate conceptual tags within the domain of expertise formalized by the ontology, expressed in OWL [8]. The result of the annotation process is a set of instances that, together with the domain ontology, form a knowledge base. Each instance is defined by its URI, its type (i.e. the class it belongs to) and by attribute values and relations that have been specified/computed. In its current version, the ShapeAnnotator saves the multi-segmented mesh along with the selected features as a single PLY file. The instances are saved as a separate OWL file that imports the domain ontology. Additionally, the OWL file contains the definition of two extra properties:

- `ShannGeoContextURI`, whose value is the URI of the multi-segmented mesh (typically the path to the PLY file saved by the ShapeAnnotator);
- `ShannSegmentID`, whose value is an index that specifies a segment in the multi-segmented mesh.

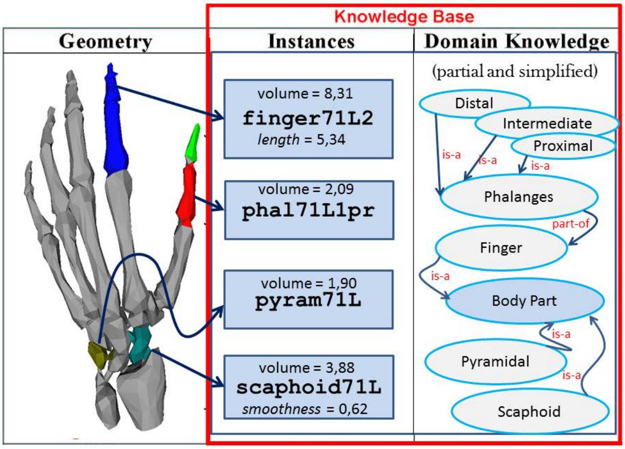


Fig. 1. An example of coupling geometric and semantic information on a human hand

All the instances produced during the annotation pipeline are automatically assigned values for the above two properties, so that the link between semantics and geometry is maintained within the resulting knowledge base (see Fig. 1). In this way, every component of the 3D model has its unique reference to its geometry and its descriptive tag, which constitute a first step towards an intelligent 3D indexing in a multi-modal knowledge based system.

Currently, the system requires the user to select manually the concepts to instantiate; for some attributes and relations, however, there is the possibility to calculate them automatically without the user intervention. Concepts of the input ontology may be equipped with descriptive attributes; for instance, a part annotated as *bone* may be described by its *volume*, *area*, and *length* as well as by other more complex attributes like *compactness*, *roundness*, or *smoothness*. How to compute the values of these attributes?

The ShapeAnnotator comes with a set of tools able to compute a number of geometric measures of shape parts (e.g. bounding box length, radius of best-fitting cylinder). These measures can be connected by the user to the attributes of the ontology in order to assign them a geometric interpretation. This allows the system to compute and fill in the values to be associated to the attributes automatically. For instance, the attribute "`length`" can be connected to, i.e. interpreted as, the *length of the first principal component* and its values computed automatically. The same kind of connections may be established to give specific interpretations to relationships. For instance, the conceptual relation "`is_connected_to`" can be connected to *topological adjacency* between the part boundaries.

Users are free to set up their "interpretation" within each specific domain of annotation by establishing an appropriate set of *connections*. Connections create a bridge between the geometrical world and the conceptual world.

Part-based annotations are important also to support comparative analysis of clinical cases: indeed, descriptions attached to parts can be used as actual *signatures*. Signatures are abstract descriptions of the content of the original resource and allow comparisons and similarity assessment [9,10]; in the medical domain they are helpful to automate part classification, to ease 3D part-based retrieval, and to monitor the changes of a specific part over time. Vast surveys on retrieval issues can be found in [11,12].

Just to name some properties that can be extracted from 3D models and used as signatures, *Shape Distributions* by Osada et al. [13] measure the distribution of properties based on distance, angle, area, and volume measurements between random surface points; Zhang and Chen [14] propose methods to compute efficiently global measures such as volume, area, statistical moments. Finally, the *configuration* of the shape features may be also detected to perform some kind of structural similarity assessment: various techniques exist to produce such signature, which range from the use of *skeletonization* to topological approaches [15,16].

3 A Medical Scenario: The Case of Rheumatoid Arthritis

The potential of part-based annotation is well demonstrated in applied cases, among which we have selected the early diagnosis and the follow-up of Rheumatoid Arthritis (RA). In the following, we will discuss the issues that motivate the adoption of 3D characterizations of parts, contextualizing the discussion to the functionalities and purpose of the *RheumaSCORE* software.

RA is a systemic, inflammatory disease that affects the synovial joints and leads to joint pain, stiffness and limited motion. It is a chronic disease that affects about 2,9 million people in Europe [17,18]. An early diagnosis, the continuous monitoring of disease activity and the constant evaluation of therapy effects can improve patients' quality of life and may reduce related social costs.

In order to evaluate RA progression and joint damage, a lot of laboratory tests are available, such as Rheumatoid Factor, C-Reactive Protein and instrumental exams, such as Magnetic Resonance Imaging. MRI has been demonstrated to be from two to ten times more sensitive than conventional radiography in detecting wrist erosions in RA (Fig. 2), especially in its early phases [19]. In general, erosions detectable on MRI may become visible on plain x-rays only 2-6 years later [20-22]. This increased sensitivity is explained by the fact that MRI is a multi-planar technique. Moreover, it can image the soft tissues, including synovial membrane, synovial fluid and tendons, in addition to bones and cartilage. The quantification of synovial volume can be used to monitor the response to therapy and to predict which patients are more likely to develop erosions within one year [23].

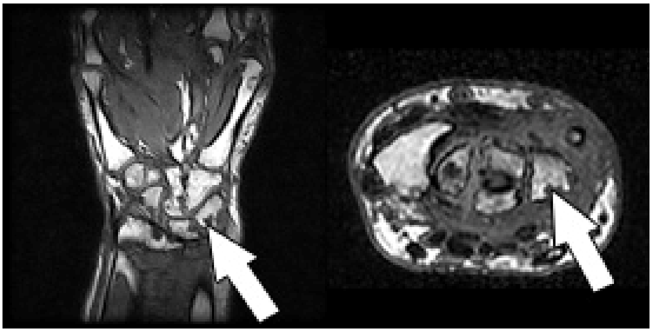


Fig. 2. MRI erosions

The wide use of MRI in the assessment of joints of patients with RA in the last years emphasized the need for an objective and reproducible scoring system of RA lesions. An international Outcome Measures in Rheumatology Clinical Trials (OMERACT) MRI in RA working group developed a MRI scoring system to assess both inflammation (activity) and bone lesions (damage) [24].

The OMERACT RA-MRI Scoring system (Rheumatoid Arthritis Magnetic Resonance Image Scoring, or RAMRIS) was developed in order to measure the lesions observed in the wrist/hand of patients with RA. These lesions are the *synovitis* (inflammation of the synovial membrane of RA and other typical forms of arthritis), the *bone marrow edema* (inflammation of the bone marrow, at least in RA), and the *erosion* (the destructive bone erosion typical of RA).

The erosion score is estimated visually by the user in the traditional RAMRIS: each eroded bone is considered individually and the ratio between the volume of the erosion and the hypothetically healthy bone is evaluated, analyzing all the slices covering the bone. The global score of the erosion is evaluated considering the eroded bone volume compared to the intact bone, with 10% increments. As a result, the rating of the erosion per single bone is comprised between 0 (healthy bone) and 10. Considering all the involved bones, the total score for the wrist is between 0 and 150, and for the hand between 0 and 80.

Manual evaluation of bone erosions volume is however tedious, time consuming and not fully repeatable (especially for inexperienced users). Considering the big amount of patients suffering from RA, this is a critical task. The *RheumaSCORE* software was developed by Softeco Sismat S.r.l. to face the RA problem [25,26].

3.1 RheumaSCORE

RheumaSCORE is an easy-to-use imaging application that supports the user (e.g. radiologist or rheumatologist) during the diagnostic process and the management of RA progression, through the analysis, the display, the measurement and the comparison of MRI acquisitions of different patients.

For each patient RheumaSCORE can load several DICOM files (study/series) simultaneously, which are used to evaluate the current disease status and monitor its progress over time. The physician is supported through several functional environments addressing:

- the *investigation*, through the recognition of wrist/hand bones and the automatic evaluation of the bones erosion scoring;
- the *tracking,* through the management of clinical data (like Rheumatoid Factor and C-Reactive Protein), the insertion of free annotations and the retrieval of similar RA cases on the basis of historical clinical data, RA measurements or keywords specified in the free notes;
- the *follow-up*, through the automatic comparison among the parameter measured in image pairs acquired at different times.

The software has a modular architecture, which can be easily expanded with other segmentation techniques and 3D visualizations to deal with other anatomical districts and pathologies.

3.2 Evaluation of RA Status and Progression

RheumaSCORE allows to analyze the bones of the hand and the wrist to assess the RA status through erosion scoring and progression monitoring. The system supports the user during the 3D segmentation process of the bones structure, which is a necessary step to evaluate automatically the bone erosion scoring.

In the recognition environment the system provides a custom segmentation procedure for each element of interest (carpal, metacarpal and forearm bones): a semi-automated method based on level sets technique using Geodesic Active Contour approach [27] has been applied, which does not rely on any prior knowledge of the shape of healthy bones. Segmentation results are reconstructed in 3D using the Marching Cube algorithm [28] and displayed using surface rendering algorithms.

After segmentation, the system provides automatic scoring of the bones erosion, using the same method proposed by OMERACT RAMRIS (see Fig. 3). It identifies and measures bone erosions, defined as the missing volume of substance of the segmented bone with respect to an average statistical model, which is built on bones of healthy subjects. Processing takes a few minutes for all wrist bones (or hand bones), which leads to a remarkable reduction of diagnosis time and costs.

Moreover, the framework permits the management and storage of clinical data (like C-reactive Protein), useful for measure and monitor RA activity. Physicians can also add annotations, possibly using the system ontology, in order to highlight lessons learnt or critical issues linked to specific features of the current patient.

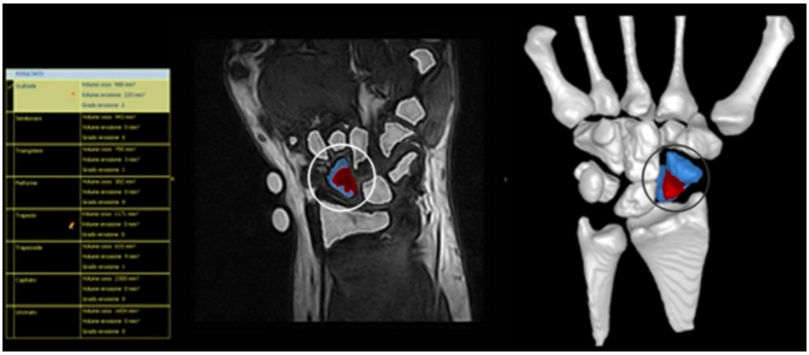


Fig. 3. Erosion Scoring

All the information related to the patient's examination (e.g. acquired DICOM images, anatomical 3D segmented elements, 3D features, user annotation) are stored in the system database and are available for retrieval.

The patient's disease follow-up is supported by storing, visualizing and comparing several sets of data acquired at different times. Differences among parameters and trends can be computed and visualized.

3.3 Clinical Trials and Results

A preliminary clinical test has been carried out at DIMI[1]. 26 patients (21 women and 5 men) diagnosed with early RA according to the 1987 ACR criteria were studied. The wrists were imaged through an extremity-dedicated MRI device (Artoscan C, Esaote, Genova Italy). A turbo T1-weighted three dimensional sequence (T3-D T1) in the coronal plane, with subsequent multiplanar reconstructions on the axial and sagittal planes, was used; slice thickness was 0,6-0,8 mm, TR 860 ms, TE 26 ms, and number of excitations (NEX) 1.

Some experts evaluated the erosion scores using the manual RAMRIS method and the RheumaSCORE software used for the segmentation of the bones of the wrist and metacarpal bases. When needed, the resulting outline was adjusted manually via a 2D editing tool. The 3D reconstruction was performed and the erosion scoring was calculated.

The median erosion scores revealed by manual application of the RAMRIS method and by automatic RheumaSCORE software were 2 (range 0-26) and 2 (range 0-21), respectively. The two scores were correlated (correlation coefficient 0.9, $p<0.0001$). The inter-rater agreement statistical measure (weighted k) was 0.706 for the entire score and was comprised between 0.264 and 0.887 for the individual bones (values greater than 0.5 are considered satisfactory). The poorest result was seen for the scaphoid due to underestimation of a relatively large erosion in a single patient (RAMRIS score 3) because of the very large size of his wrist bones. Therefore, the semi-automated segmentation software showed a good correlation with the RAMRIS erosion score, yet presenting some limitations.

[1] DIMI -Dipartimento di Medicina Interna, Clinica Reumatologica, Università degli Studi di Genova, Italy.

4 Integration Perspective

The functionalities implemented in the RheumaSCORE are a clear example of the directions that CAD systems are likely to take in the near future. Adding tools to select parts of interest in 3D reconstructions and automatically compute morphological parameters is of great value, and it has been shown to be effective for the early diagnosis of RA even in the initial stage of development. Currently, the systems uses the Geodesic Active Contour approach as the only technique for segmentation and the global volume of the detected parts as the only geometrical characteristic recorded.

The integration of this platform with the ShapeAnnotator tool and the experience in a finer-grained 3D analysis may produce very informative characterizations and indices to support classification and statistical analysis.

In the first place, it is necessary to adopt a multi-segmentation strategy: a rich library of tools for the segmentation of parts has to be provided to allow users to rely on various, and possibly integrated, segmentation techniques to locate better the interesting parts, possibly taking into account also uncertainty.

Moreover, by relying on the fine-grained 3D characterization techniques, each segment could be indexed not only by its volume, but also by a number of content-based descriptors, each of them highlighting diverse aspects of the considered part. For instance, spatial localization of bone erosion can be useful to evaluate more accurately the anatomical-functional damage, the possible disease progression, the pain felt by the patient and the effectiveness of the therapy. A new version of the CAD system would also provide morphological analysis to evaluate the bone roughness (i.e. lack of smoothness, which can be computed from the distribution of curvature on the bone surface) or the presence of interruptions of the bone cortex, which are typical of erosions. In fact, normal areas of tendon and ligament attachment could be rough because of traction forces and a very small cortical interruption can be seen where nutritional arteries enter the bone. In addition, a family of arthritis called seronegative spondyloarthritides, which enter the differential diagnosis of RA, is characterized by periostitis or inflammation of the periosteum and bone cortex. This lesion is reflected by a rough appearance of the bone surface. The new system would evaluate an increased percentage of rough bone surface and contribute to both differential diagnosis and damage follow-up in psoriatic arthritis patients.

In Fig. 4 we show some preliminary tests run on the semilunar bone, comparing the curvature plots between a healthy bone and a highly eroded one. Curvature has been computed with the algorithm developed in [29] and shows concave, convex and saddle areas. Numerical curvature values could be additionally gathered for a quantitative data analysis.

Finally, it is necessary to connect the annotation mechanism to the system in order to link each of the selected parts with concepts and attributes expressed in a given domain ontology, thus easing the documentation accuracy and the retrieval performance.

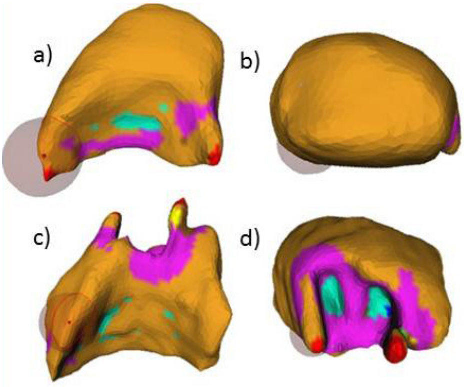


Fig. 4. Top: a) and b) are two views of a healthy semilunar bone of the wrist, which is characterized by large smooth areas. Bottom: c) and d) are two views of a RA-affected semilunar, which is well characterized by the massive presence of tips (red areas), saddles (magenta areas) and concavities (cyan areas) in the eroded zone

5 Conclusions

In this paper, we showed the high potential of including full 3D content among the heterogeneous digital data available in medicine. In fact, advanced shape analysis and similarity techniques can be exploited to improve CAD systems and content-based retrieval systems. We discussed such issues in the frame of RA, starting from the ShapeAnnotator and the RheumaSCORE tools. In particular, we mainly highlighted the promising perspectives of coupling geometric information with semantic characterization.

As a final remark, an important research challenge on the side of knowledge representation is worth to be mentioned: at the state of the art, geometric representations are not consistently linked with part-based annotations. On the contrary, it would be fundamental to define a stable markup: annotations attached to parts of 3D models should survive changes in the geometric representation (e.g. change of representation, change of resolution, shape editing). In this respect, the ShapeAnnotator provides just a partial solution and the problem of defining a stable 3D markup still remains.

Acknowledgments. This work is supported by the FP7 Marie Curie Initial Training Network "MultiScaleHuman": Multi-scale Biological Modalities for Physiological Human Articulation (2011-2015), contract MRTN-CT-2011-289897. We kindly acknowledge the support of the FP7 "VISIONAIR": Vision Advanced Infrastructure for Research (2011-2015), grant no. 262044, and of POLITECMED – Research and Innovation Pole of the Regional Centre for Research and Innovation of the Regione Liguria. Softeco wishes to thank Esaote spa for their collaboration. The RheumaSCORE software has been developed within the P.O.R. Liguria FESR (2007-2013) - Asse 1 “Innovazione e competitività”- Bando Azione 1.2.2 - Progetto SIDARMA. Finally, IMATI thanks Michela Mortara for the curvature tests on the semilunar bones.

References

1. Anibaldi, C.: Sistemi Esperti come supporto alla decisione clinica. Seminario di Reumatologia e Medicina Interna. (April 2004)
2. Leitich, H., Adlassing, K.P., Kolarz, G.: Evaluation of two different models of semi-automatic knowledge acquisition for the medical consultant system CADIAG-II/RHEUMA. Artificial Intelligence in Medicine (October 2001)
3. Aisen, M., Broderick, L.S., Winer-Muram, H.C., Brodley, E., Kak, A.C., Pavlopoulou, C., Dy, J., Shyu, C.R., Marchiori, A.: Automated Storage and Retrieval of Thin-section CT Images to Assist Diagnosis. System Description and Preliminary Assessment. Radiology 228(1), 265–270 (2003)
4. Attene, M., Biasotti, S., Mortara, M., Patanè, G., Spagnuolo, M., Falcidieno, B.: Computational methods for understanding 3D shapes. Computers & Graphics-Uk, Special Issue on Computer Graphics in Italy 30(3), 323–333 (2006)
5. Catalano, C.E., Mortara, M., Spagnuolo, M., Falcidieno, B.: Semantics and 3D media: Current issues and perspectives. Computers & Graphics 35(4), 869–877 (2011)
6. Attene, M., Robbiano, F., Spagnuolo, M., Falcidieno, B.: Characterization of 3D Shape Parts for Semantic Annotation. Journal of Computer Aided Design 41(10), 756–763 (2009)
7. Shamir, A.: A Survey on Mesh Segmentation Techniques. Computer Graphics Forum 27(6), 1539–1556 (2008)
8. OWL web ontology language guide, w3C, `http://www.w3.org/TR/owl-guide/`
9. van Kaick, O., Zhang, H., Hamarneh, G., Cohen-Or, D.: A Survey on Shape Correspondence. Computer Graphics Forum (CGF) (Extended Version of Eurographics 2010 STAR) 30(6), 1681–1707 (2011)
10. Bustos, B., Keim, D., Saupe, D., Schreck, T.: Content-Based 3D Object Retrieval. IEEE Comput. Graph. Appl. 27(4), 22–27 (2007)
11. Tangelder, J.W.H., Veltkamp, R.C.: A survey of content-based 3D shape retrieval methods. Multimedia Tools and Applications 39(3), 441–471 (2008)
12. Yang, Y., Lin, H., Zhang, Y.: Content-based 3D model retrieval: a survey. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 37(6), 1081–1098 (2007)
13. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distribution. ACM Trans. Graph. 21(4), 807–832 (2002)
14. Zhang, C., Chen, T.: Efficient feature extraction for 2D/3D objects in mesh representation. In: IEEE International Conference on Image Processing. IEEE Computer Society Press (2001)
15. Sundar, H., Silver, D., Gagvani, N., Dickinson, S.: Skeleton Based Shape Matching and Retrieval. In: International Conference on Shape Modeling and Applications, pp. 130–139 (2003)
16. Biasotti, S., Giorgi, D., Spagnuolo, M., Falcidieno, B.: Reeb graphs for shape analysis and applications. Theoretical Computer Science 392(1-3), 5–22 (2008)
17. Markenson, J.A.: Worldwide trends in the socioeconomic impact and long-term prognosis of rheumatoid arthritis. Semin. Arthritis Rheum. 21, 4–12 (1991)
18. Weinblatt, M.E.: Rheumatoid arthritis: treat now, not later (editorial). Ann. Intern. Med. 124, 773–774 (1996)
19. Østergaard, M., Hansen, M., Stoltenberg, M., Jensen, K.E., Szkudlarek, M., Pedersen-Zbinden, B., Lorenzen, I.: New radiographic bone erosions in the wrists of patients with rheumatoid arthritis are detectable with magnetic resonance imaging a median of two years earlier. Arthritis Rheum. 48, 2128–2131 (2003)

20. Benton, N., Stewart, N., Crabbe, J., Robinson, E., Yeoman, S., McQueen, F.M.: MRI of the wrist in early rheumatoid arthritis can be used to predict functional outcome at 6 years. Ann. Rheum. Dis. 63, 555–561 (2004)
21. McQueen, F.M., Benton, N., Perry, D., Crabbe, J., Robinson, E., Yeoman, S., McLean, L., Stewart, N.: Bone edema scored on magnetic resonance imaging scans of the dominant carpus at presentation predicts radiologic joint damage of the hands and feet six years later in patients with rheumatoid arthritis. Arthritis Rheum. 48, 1814–1827 (2003)
22. Østergaard, M., Hansen, M., Stoltenberg, M., Jensen, K.E., Szkudlarek, M., Klarlund, M., Pedersen-Zbinden, M.: MRI bone erosions in radiographically non-eroded rheumatoid arthritis wrist joint bones give a 4-fold increased risk of radiographic erosions five years later. Arthritis Rheum. 46(suppl.), S526–S527 (2002)
23. Savnik, A., Malmskov, H., Thomsen, H.S., Graff, L.B., Nielsen, H., Danneskiold-Samsøe, B., Boesen, J., Bliddal, H.: MRI of the wrist and finger joints in inflammatory joint diseases at 1-year interval: MRI features to predict bone erosions. Eur. Radiol. 12, 1203–1210 (2002)
24. Ejbjerg, B., McQueen, F., Lassere, M., Haavardsholm, E., Conaghan, P., O'Connor, P., Bird, P., Peterfy, C., Edmonds, J., Szkudlarek, M., Genant, H., Emery, P., Ostergaard, M.: The EULAR-OMERACT rheumatoid arthritis MRI reference image atlas: the wrist joint. Ann. Rheum. Dis. 64(suppl. 1), 23–47 (2005)
25. RheumaSCORE, `http://www.research.softeco.it/rheumascore.aspx`
26. Barbieri, F., Parascandolo, P., Vosilla, L., Cesario, L., Viano, G., Cimmino, M.A.: Assessing MRI erosions in the rheumatoid wrist: a comparison between RAMRIS and a semiautomated segmentation software. Ann. Rheum. Dis. 71(suppl. 3) (2012)
27. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. International Journal of Computer Vision 22(1), 61–79 (1997)
28. Lorensen, W.E., Cline, H.E.: Marching Cubes: a high resolution 3D surface construction algorithm. ACM SIGGRAPH, 163–169 (1987)
29. Mortara, M., Patanè, G., Spagnuolo, M., Falcidieno, B., Rossignac, J.: Blowing Bubbles for Multi-Scale Analysis and Decomposition of Triangle Meshes. Algorithmica 38(1), 227–248 (2004)

Skull Retrieval for Craniosynostosis Using Sparse Logistic Regression Models

Shulin Yang[1], Linda Shapiro[1], Michael Cunningham[2], Matthew Speltz[2],
Craig Birgfeld[2], Indriyati Atmosukarto[3], and Su-In Lee[1]

[1] Computer Science and Engineering, University of Washington, Seattle, WA
{yang,shaprio,suinlee}@cs.washington.edu
[2] Seattle Children's Research Institute, Seattle, WA
{michael.cunningham,matt.speltz,craig.birgfeld}@seattlechildrens.org
[3] Advanced Digital Sciences Center, Singapore
indria@adsc.com.sg

Abstract. Craniosynostosis is the premature fusion of the bones of the calvaria resulting in abnormal skull shapes that can be associated with increased intracranial pressure. While craniosynostoses of multiple different types can be easily diagnosed, quantifying the severity of the abnormality is much more subjective and not a standard part of clinical practice. For this purpose we have developed a severity-based retrieval system that uses a logistic regression approach to quantify the severity of the abnormality of each of three types of craniosynostoses. We compare several different sparse feature selection techniques: L_1 regularized logistic regression, fused lasso, and clustering lasso (cLasso). We evaluate our methodology in three ways: 1) for classification of normal vs. abnormal skulls, 2) for comparing pre-operative to post-operative skulls, and 3) for retrieving skulls in order of abnormality severity as compared with the ordering of a craniofacial expert.

Keywords: craniosynostosis, cranial image (CI), L_1 penalized logistic regression, fused lasso, clustering lasso (cLasso), sparse logistic regression model.

1 Introduction and Motivation

This work is focused on retrieval of CT images for patients with craniosynostosis, a common congenital condition in which one or more of the fibrous sutures in an infant's calvaria fuse prematurely, resulting in restricted skull and brain growth. Because the brain cannot expand perpendicular to the fused suture, it redirects growth in the direction of the open sutures, resulting in abnormal head shape and in some cases, facial features. Craniosynostosis results in head deformity that can be severe if it is not corrected surgically. This condition may result in increased intracranial pressure on the brain and is correlated with developmental delays, although the cause of such delays is not currently known [11]. It is estimated that the fusion of any one or more sutures occurs in approximately 1 in 2,000

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 33–44, 2013.

live births [7]. In clinical practice, craniosynostosis is diagnosed by a physician on the basis of head shape and confirmatory CT scan.

Automatic analysis of CT scans, including a measure of shape deformation, would be of great help to both doctors and medical researchers. In our previous work, we built a system that automatically generates a shape representation called the cranial image (CI) [4] from the CT image of a patient's skull. The cranial images are used as shape features to distinguish between skulls of patients with different types of craniosynostosis. We also proposed using logistic regression and three variations of the logistic regression model for classifying different types of craniosynostosis: L_1 regularized logistic regression [2], the fused lasso [9] and the clustering lasso (cLasso) [1], which is a variation of L_1 logistic regression. These models avoid overfitting of the regression model, and they also could select subsets of features from the cranial image that represent skull regions associated with the distinctive shape differences related to different suture fusions (e.g., sagittal vs metopic suture fusion).

It is important to note that clinicians do not need to rely on a quantitative model to make the diagnosis of craniosynostosis. However, there is a lack of criteria to quantify the severity of the abnormality for research purposes. For example, when estimating the relative effects of different surgical methods on cranial shape (i.e., pre-, post-surgery change), quantitative measurement is essential. For this reason, we have developed a system to retrieve CT images based on quantification of the severity of the abnormality of the 3D skull shape. Given an enlarged data set containing pre-operative and post-operative CT scans of subjects with three classes of craniosynostosis (coronal, metopic and sagittal) plus a set of scans from similar-age control subjects, we conducted a set of experiments in classification, quantification and retrieval using the three logistic regression methods proposed in [1]. Different sparse logistic regression models are compared in terms of misclassification on whether a skull has craniosynostosis or not. Then we show our retrieval results using the best model for our data - cLasso. Abnormality of the skulls of the same patient before a surgery and after a surgery is compared using the quantification criteria as well.

The rest of the paper is organized as follows. Section 2 summarizes the related literature, Section 3 gives an overview of the framework of our approach for abnormality quantification, Section 4 describes the details on how logistic regression models are used for quantification, and Section 5 shows the experimental results of our work.

2 Related Literature

Calvarial (skull) abnormalities are frequently associated with severely impaired central nervous system functions due to brain abnormalities, increased intracranial pressure and abnormal build-up of cerebrospinal fluid. In [3], Shapiro et al. introduced several different craniofacial descriptors that have been used in studies of two craniofacial disorders: 22q11.2 deletion syndrome (a genetic disorder) and deformational plagiocephaly/brachycephaly. They provided feature extraction tools for the study of craniofacial anatomy from 3D mesh data obtained from the 3dMD active stereo photogrammetry system. These tools produce quantitative representations (descriptors) of the 3D data that can be used to summarize the 3D shape as pertains to the condition being studied and the question being asked. This work is different from the current study in that it analyzed the shape of the midface and back of the head, while our work focuses on the shape of the skull.

There are some previous studies on examining the specific skull shapes of patients. Previously, we proposed the cluster lasso [1], a logistic regression model, for classification of three types of craniosynostoses: coronal, metopic and sagittal. Lin et al. [4] developed symbolic shape descriptors to classify skull deformities caused by metopic and sagittal synostoses. Ruiz-Correa et al. [5] used a set of scaphocephaly severity indices (SSIs) for predicting and quantifying head- and skull-shape deformity in children diagnosed with isolated sagittal synostosis (ISS).

The study differs from previous work in that it focuses on the task of skull retrieval, while our previous work focused on classifying different types of craniosynostoses. Another difference is that the current work fits a regression model based on abnormal shapes versus normal ones, while in our previous work we were merely comparing different types of abnormal shape and could therefore not quantify severity with respect to normal. Other efforts mentioned above differ from our approach in that they were not fully automatic and therefore required human interaction for selecting planes and landmarks from the skull for the purpose of extracting shape features.

3 Skull Retrieval Pipeline

3.1 System Design

A system was built for skull retrieval according to shape abnormality severity. With an input of 3D CT volume data of random pose, our system first extracts the skull and performs pose normalization, so that it is symmetric with respect to the right and left sides. Then, surface points are extracted that are evenly spaced all over the skull. After that, a shape feature called the cranial image [4] is calculated by computing pairwise distances of these points. Last, a method is proposed to quantify skull abnormality severity using the shape feature.

3.2 Surface Points Extraction

The first step of this module is to locate a base plane on the skull based on two important landmarks: the nasion and the opisthion. The base plane goes through these two biological landmarks, and is perpendicular to the symmetry plane that separates the right and left sides.

The nasion is the intersection of the frontal and two nasal bones of the human skull [8]. Its manifestation on the visible surface of the face is a distinctly depressed area directly between the eyes, just superior to the bridge of the nose.

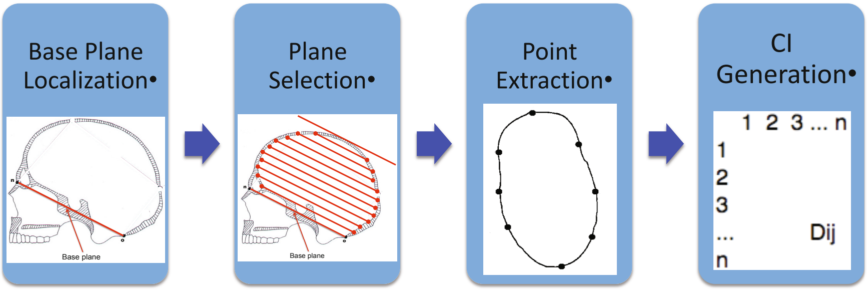


Fig. 1. Surface points extraction: the first three modules are the three steps for extracting surface points; the last image is the distance matrix generated from the surface points

The opisthion is the mid-point of the posterior margin of the foramen magnum on the occipital bone [8]. The two points were chosen because of their locations at the front and back of the head, and because they are stable during the human growth process The nasion and opisthion are detected as follows. First, the plane of symmetry of the left and right sides of the skull on which the landmarks are expected to be is extracted. Then, the tip of the nose is located as the point with the smallest horizontal value. The nasion is located as the point closest to the tip of the nose, which is above it and which has a zero curvature in the vertical direction. The opisthion is located as the point in the left part of the outline, which has the closest distance to the nasion of all points below the nasion.

Our shape measure is based on the distances between points on the surface of the skull, so the second step is to extract a set of planes from which surface points are located. These planes are parallel to the base plane. The top plane of the skull is a plane that has intersection with the skull and which is parallel to the base plane but has the furthest distance to the base plane. Our system can extract any plane that is parallel to the base plane and located between the base plane and the top plane of the skull, based on the ratio of its distance to these two planes. Multiple planes may be selected with equal distances among them. In the rest of our experiments, 10 planes that are evenly distributed across the whole skull were used to provide a rich 3D shape descriptor.

In the third step of this module, N points are evenly extracted along the outlines of the planes from the previous step. N is chosen by the user, and $N = 100$ in our experiments.

3.3 Cranial Image Generation

Our shape feature is a $N \times N$ pairwise distance matrix among the surface points from the previous step. The number at position (i, j) of the matrix represents the distance between point number i and point number j (the last module of in Fig. 1). The matrix is symmetric. Such a shape feature is a rich representation of the skull shape, and its dimension is usually very high (over 10^4).

3.4 Abnormality Quantification

The task of the system is to retrieve CT images by their skull shape abnormality severity. To rank the CT images, a qualification criterion for shape abnormality is needed. Based on our previous work [1], sparse logistic regression models have been effectively used to fit our data with the high dimensional shape features (CI). We use these models to generate quantification scores that represent the severity of the abnormality.

4 Sparse Logistic Regression Models for Abnormality Quantification

4.1 Sparse Logistic Regression Models

In our previous work, we explored logistic regression and three sparse logistic regression models for the purpose of classifying three different types of craniosynostosis. They can be summarized as follows.

Logistic Regression. Logistic regression is a workhorse in machine learning that uses a generalized linear model for binomial regression. In logistic regression model, the probability of being classified as one class is a linear function of the features.

$$p(y|\mathbf{x},\mathbf{w}) = \frac{1}{1 + exp(-y(\mathbf{w}^T\mathbf{x} + w_0))} \tag{1}$$

where vector $\mathbf{x}$ contains the feature values of a data sample; y is its class label (for example, $y = 1$ refers to coronal and $y = -1$ refers to non-coronal), $\mathbf{w}$ contains the coefficients for $\mathbf{x}$, and w_0 is the intercept. Furthermore, w_0 and $\mathbf{w}$ are model parameters, and $p(y|\mathbf{x},\mathbf{w})$ is the probability that a data sample belongs to a certain class.

We can estimate the optimal parameters w_0 and $\mathbf{w}$ that minimize the following loss function:

$$l(w_0,\mathbf{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x_i} + w_0))) \tag{2}$$

$$\{w_0,\mathbf{w}\} = \min_{w_0,\mathbf{w}} l(w_0,\mathbf{w}) \tag{3}$$

where y_i is the actual class label of a data sample $\mathbf{x_i}$.

$\mathbf{L_1}$ Regularized Logistic Regression. Due to the high-dimensionality of the data (i.e. a large number of features and a modest size of samples), learning the unregularized logistic regression [3] will result in overfitting. To avoid overfitting, L_1 regularization is usually applied to induce sparsity in the solution $\mathbf{w}$ such that many of the coefficients in $\mathbf{w}$ are set to exactly zero. L_1 regularization [2] has been rigorously proven to be effective in selecting relevant features when there are exponentially many irrelevant ones [6]. The log-likelihood of L_1 regularized logistic regression is as follows.

$$l(w_0, \mathbf{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T \mathbf{x_i} + w_0))) + \lambda \sum_{i=1}^{m} |w_i| \tag{4}$$

where λ is a regularization parameter for the L_1-norm of the coefficients.

Fused Lasso. One problem with L_1 regularization is that when features are highly correlated, it arbitrarily chooses one of many correlated features. Some variations of L_1 regularization can better exploit the underlying structure of our feature data. Specifically, the fused lasso induces bias from prior knowledge such that correlated feature groups will be assigned similar weights. In this work, the fused lasso [9] places a constraint on the weights of the features that are geographically related - sharing the same or neighboring surface points.

The loss function of the fused lasso with induced bias is,

$$\begin{aligned} l(w_0, \mathbf{w}) = &\sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T \mathbf{x_i} + w_0))) \\ &+\lambda \sum_{i=1}^{m} |w_i| + \mu \sum_{\{w_i, w_j\} \in M} |w_i - w_j| \end{aligned} \tag{5}$$

where μ is a regularization parameter for the new penalty term. M is a set that contains all pairs of features that are neighbors, whose endpoints are the same or next to each other. In equation [5], $\lambda \sum_{i=1}^{m} |w_i|$ penalizes large feature weights, and $\mu \sum_{\{w_i, w_j\} \in M} |w_i - w_j|$ penalizes large weight differences between correlated features.

Clustering Lasso. As mentioned before, L_1 regularized logistic regression tends to assign different weights to highly correlated features. When features are highly correlated, it arbitrarily chooses one of them and assigns a non-zero weight only to it. The fused lasso is one way to avoid this problem by placing constraints on the weight differences based on prior knowledge. However, this requires the model to know ahead of time the right grouping of the features. An alternative to using such prior knowledge, is to penalize the weight differences of correlated features.

The clustering lasso (or cLasso) is a new form of regularized logistic regression we recently proposed in [1]. The model for the clustering lasso is:

$$p(y|\mathbf{x}, \mathbf{w}, \mathbf{w^c}) = \frac{1}{1 + exp(-y(\mathbf{w}^T \mathbf{x} + \mathbf{w^c}^T \mathbf{c} + w_0))} \tag{6}$$

where $\mathbf{x}$ contains the feature values of a data sample; y is its class label (for example, $y = 1$ refers to sagittal and $y = -1$ refers to non-sagittal); $\mathbf{c}$ are the cluster centers of $\mathbf{x}$; $\mathbf{w}$ contains the coefficients for $\mathbf{x}$; $\mathbf{w^c}$ contains the coefficients for $\mathbf{c}$; and w_0 is the intercept. Furthermore, w_0, $\mathbf{w}$ and $\mathbf{w^c}$ are model parameters, while $p(y|\mathbf{x}, \mathbf{w}, \mathbf{w^c})$ is the probability that a data sample belongs to a certain class.

The loss function for the cLasso is

$$l(w_0, \mathbf{w}, \mathbf{w^c}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x_i} + \mathbf{w^c}^T\mathbf{c_i} + w_0))) \\ +\lambda \sum_{i=1}^{m} |w_i| + \nu \sum_{i=1}^{k} |w_i^c| \qquad (7)$$

where $\mathbf{c_i}$ $(i \in [1, k])$ is the centroid of a group of features $\{x_{i_1},x_{i_2},...,x_{i_k}\}$ (its feature value is their average); w_i^c is the weight for c_i; and ν is the regularization parameter for the weights of the cluster centers.

This loss function is designed to cluster the features based on their correlation, and penalize their shared weights (w_i^c) and individual weights $(w_{i_1}, w_{i_2}, ..., w_{i_k})$ respectively. When ν is small and λ is large, individual weights are penalized, and features tend to be split into groups based on their correlation and to share the same weights. When λ is large enough, this model is equivalent to the model of L_1 regularized logistic regression (equation [4]).

Parameter w_i^c encourages correlated features to share the same weight, and w_i allows unique features to be used. Therefore, the cLasso is equivalent to using only shared weights (w_i^c) when each centroid c_i $(i \in [1, k])$ is computed as a weighted average with the weights determined by $(w_{i_1}, w_{i_2}, \ldots, w_{i_k})$.

4.2 Abnormality Quantification

Using the models to fit to the data, the predicted probability of a sample data set $\mathbf{x}$ being a certain class $p(y = 1|\mathbf{x}, \mathbf{w})$ can be viewed as a quantification measure of skull abnormality. However, the use of the sigmoid function $P(t) = \frac{1}{1+e^{-t}}$ in computing the probability does not work well as a quantification criteria. Because a regression model that fits the data well tends to assign a value close to 1 to all positive instances, the quantification results are too similar. Instead, we use the linear function of the features before taking the sigmoid function to obtain the probability. The second option produces a better quantification measure, because the linear function differentiates abnormal skulls better, even when they are classified as the same class.

For the logistic regression, lasso and fused lasso models, the quantification scores are

$$S(\mathbf{x}) = -y(\mathbf{w}^T\mathbf{x} + w_0) \qquad (8)$$

For the clustering Lasso, the quantification score is

$$S(\mathbf{x}) = -y(\mathbf{w}^T\mathbf{x} + \mathbf{w^c}^T + w_0) \qquad (9)$$

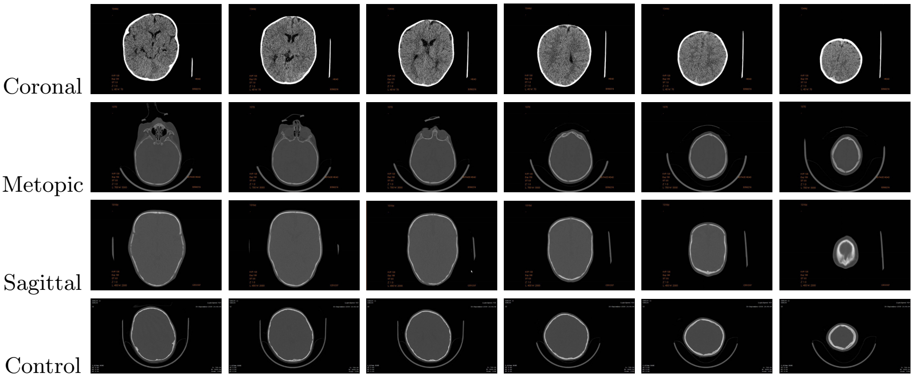


Fig. 2. Examples of CT image slices

5 Experiments

The experiments were designed to show the ability of our system to retrieve CT images based on skull shape abnormality of different types of craniosynostosis. As a measurement for performance, we also provide phenotype prediction accuracy by comparing our prediction with the groundtruth diagnoses of the doctors. Cranial images were generated using our system with 10 planes and 10 points on each plane. Logistic regression, L_1 regularized logistic regression, the fused lasso and the cLasso were compared in terms of phenotype prediction accuracy, while the quantification measure of cLasso is used as the ranking criterion in our system. Implementation of L_1 regularized logistic regression is from the authors of [6], and implementation of the fused lasso is the machine learning package SLEP [10]. In our previous work, we did a thorough study on choosing regularization parameter λ, μ and ν in equations 4, 5 and 7 for the sparse logistic regression models. We continue to use these parameters.

5.1 Medical Data

Our system was tested on 3D CT images of children's heads from hospitals in four different cities in the US. There are different types of craniosynostosis depending on the affected suture; the sagittal suture is between the parietal bones, metopic between the frontal bones, coronal between the frontal and parietal bones, and lambdoid between the parietal and occipital bones. Our study is focused on three types of synostosis - sagittal, metopic, and coronal. In total we examined approximately 200 CT image volumes, each comprising a stack of image slices (approximately 150 slices per volume). About half the data are controls (normal skulls), and the other half have one of the three types of craniosynostosis. Fig. 2 shows some examples of slices from the CT image stacks for all four classes.

5.2 Evaluation on Classification

In the first part of our experiment, we evaluate our approach based on its prediction of whether a skull is abnormal or not, meaning the misclassification rate of each class. Classification results using the four different models are shown in Table 1. The results of logistic regression substantiates the overfitting problem with the large number of features it uses. The misclassification rate greatly improves when regularization is used in the logistic regression model. Specifically, the clustering lasso exhibits the best results on average. The misclassification rate of the coronal class is higher than the other two classes. This is because the coronal class is the most similar to the controls of all three classes. Its shape deformation is not immediately obvious as the deformation of the sagittal and metopic classes, as can be observed from Figure 2. This is consistent with the results from [1].

Table 1. Misclassification rates using multiple planes for three types of craniasynostosis versus controlled skulls. Four different logistic regression models are used for comparison. Classo performs the best on all three types.

Misclassification Rate	Coronal vs Control	Metopic vs Control	Sagittal vs Control
Logistic regression	37.5%	36.25%	30%
L_1 regression	26.3%	8.75%	**8.75%**
Fused lasso	16.3%	21.3%	**8.75%**
Clustering lasso	**14.1%**	**7.5%**	**8.75%**

5.3 Abnormality Quantification of Pre- and Post-Operative Skulls

Besides testing our quantification measure on the CT images of patients when they are diagnosed, we also tested it on their CT scans two years after skull surgery was performed. Although the skull abnormality is corrected with surgery,

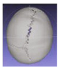
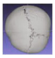
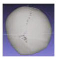
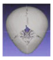
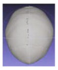
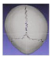
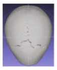
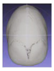
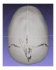
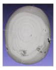
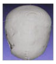
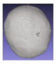
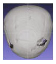
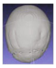
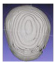
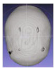
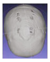
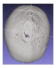
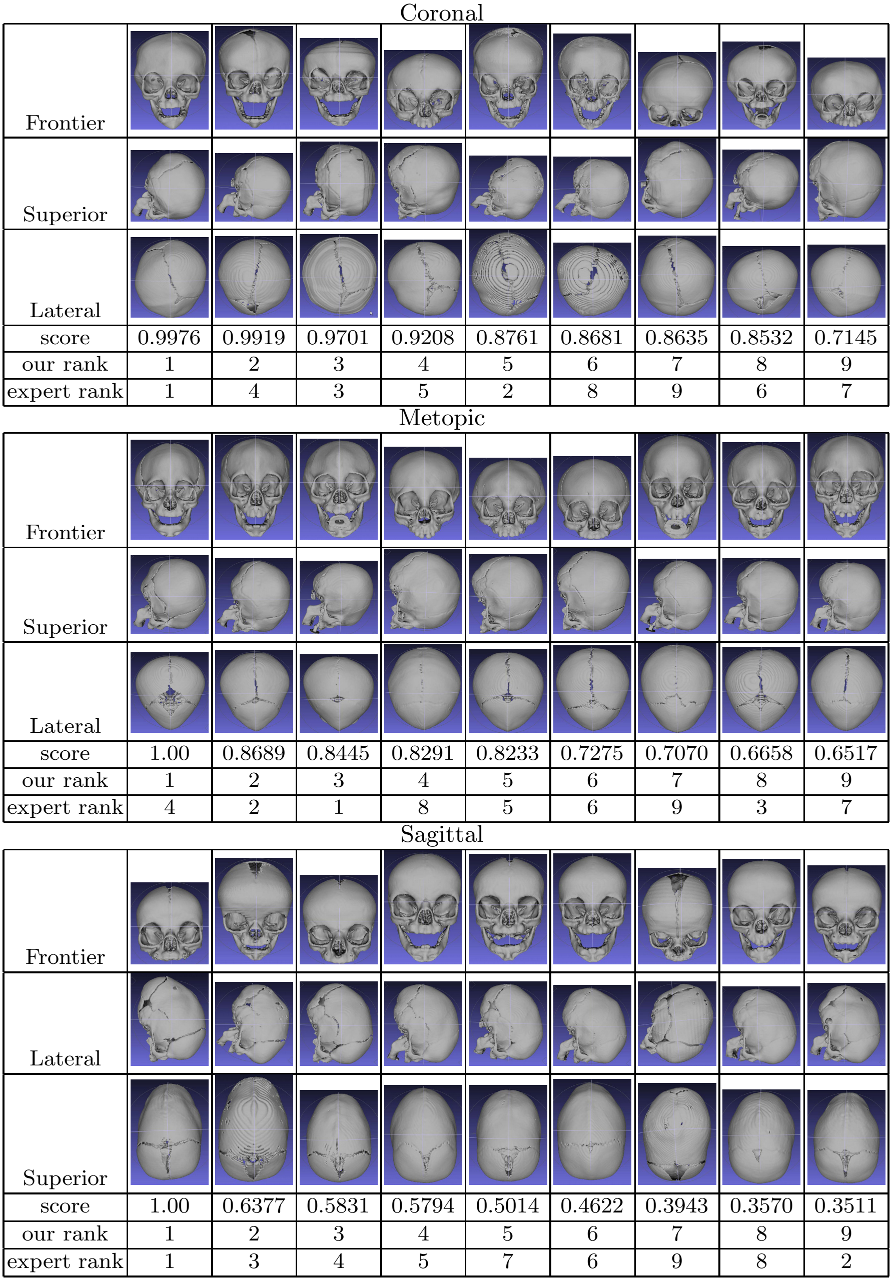
type	coronal	coronal	coronal	metopic	metopic	metopic	sagittal	sagittal	sagittal
Pre-Op	 0.9698	 0.9208	 0.7145	 1.00	 0.5355	 0.9773	 0.3464	 0.5794	 0.5831
Post-Op	 0.4164	 0.3775	 0.3930	 0.5471	 0.3134	 0.5471	 0.1890	 0.3745	 0.4895
ratio	43%	41%	55%	55%	59%	56%	55%	65%	84%

Fig. 3. Quantification comparison of pre-operative and post-operative skulls: these images are the superior views of nine skulls (for each symptom type). Under the images are their severity scores. The upper row contains pre-operative skulls, and the bottom row are the post-operative skulls of the same subjects as the pre-op ones above it. A subject is normal if the score is ≤ 0. The larger a number is, the more abnormal it is.



Coronal

Frontier									
Superior									
Lateral									
score	0.9976	0.9919	0.9701	0.9208	0.8761	0.8681	0.8635	0.8532	0.7145
our rank	1	2	3	4	5	6	7	8	9
expert rank	1	4	3	5	2	8	9	6	7

Metopic

Frontier									
Superior									
Lateral									
score	1.00	0.8689	0.8445	0.8291	0.8233	0.7275	0.7070	0.6658	0.6517
our rank	1	2	3	4	5	6	7	8	9
expert rank	4	2	1	8	5	6	9	3	7

Sagittal

Frontier									
Lateral									
Superior									
score	1.00	0.6377	0.5831	0.5794	0.5014	0.4622	0.3943	0.3570	0.3511
our rank	1	2	3	4	5	6	7	8	9
expert rank	1	3	4	5	7	6	9	8	2

Fig. 4. Quantification results: nine skulls are shown for each type of synostosis (coronal, metopic, and sagittal) from three different views. They are ordered by the severity scores produced by our system for the craniosynostosis type to which they belong. Under the images are their severity scores, their ranks by our system (ordered 1-9, with 1 being most severe), and expert ranks for comparison.

it tends to relapse toward the original deformity with time. Our quantitative severity measure provides an objective measure for the comparison. Fig. 3 shows the comparison results for a set of pre-operative and post-operative skulls. The pair of skulls in each column are from the same patient. The abnormality reduction after surgery is according to our scoring method. This shows that the surgeries resulted in improvement in all cases even after two years of growth.

5.4 Evaluation of Skull Retrieval

There is no gold standard for evaluating the quantification results, because an objective judgement of severity of craniosynostosis in the form of a medical test does not exist. Medical experts are accustomed to providing diagnoses but have no scoring criteria. In fact, our work was motivated by the need for severity quantification in medical research. However, we were able to have a craniofacial expert rank a subset of the skulls for each type of abnormality for our comparisons.

Based on the quantification measure using the clustering lasso, each skull was assigned a value that represents the degree of severity of the craniosynostosis type to which it belongs (coronal, metopic or sagittal). The results for these skulls in each class are shown in Fig. 4. (The numbers are normalized from -1 to 1; 0 to 1 means a subject is abnormal, and -1 to 0 means a subject is a control.) This result provides a useful measurement for physicians and researchers to quantify the severity of individual cases with different types of craniosynostosis. Our results correlate well with expert measures. Most of the orderings are only slightly permuted from the expert's ranking. This is because some of the severity scores are very similar in themselves, so any of them can be ranked before the others in a subjective ordering. There are several exceptions on which our system and the expert disagree in an obvious way, such as No. 4 in the metopic group and No. 9 in the sagittal group in Fig. 4. Discussion with the expert disclosed that different parts of the skull were being weighted differently by the system and the expert. For example for No. 4 in the metopic group, the expert ranked it last because his focus, the pointy shape at the front, is not obvious in the skull. However, our system ranked it higher because its global shape is more similar to a typical metopic skull. This discovery reversely inspired the expert to notice certain shape features that he had previously ignored. The expert also told us that our quantification captured some shape features that could not be observed from the three standard views by doctors. This is evidence that our system would be helpful for clinicians to make more accurate diagnoses.

6 Conclusions

In this work, we built a system that performed skull analysis and severity based retrieval for patients with craniosynostosis. The system was tested with four different logistic regression models: logistic regression, L_1 regularized logistic regression, the fused lasso and the clustering lasso on classifying abnormal skulls from normal ones. The cLasso model was used for quantification of skull shape abnormality. Our experimental results validate the model, both by the error rate on the classification task and by comparison with expert ranks on the retrieval task. This retrieval system provides a convenient tool that can help medical researchers to quantify craniosynostosis for research studies. For example, the methods reported here would facilitate studies of the effects of different surgical methods on cranial shape, associations between severity of cranial deformation and subsequent neurodevelopmental outcomes and the relation of severity to genetic processes.

Acknowledgments. This research was supported by NIH/NIDCR under grant number U01 DE 020050 and grant number R01 DE 13813, and by the Jean Renny Endowment for Craniofacial Medicine.

References

1. Yang, S., Shapiro, L., Cunningham, M., Speltz, M., Lee, S.-I.: Classification and Feature Selection for Craniosynostosis. In: Proceeding BCB 2011 Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, pp. 340–344 (2011)
2. Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society 58(1), 267–288 (1996)
3. Shapiro, L., Wilamowska, K., Atmosukarto, I., Wu, J., Heike, C., Speltz, M., Cunningham, M.: Shape- Based Classification of 3D Head Data. In: ICIAP, pp. 692–700 (2009)
4. Lin, H., Ruiz-Correa, S., Sze, R., Cunningham, M., Speltz, M., Hing, A., Shapiro, L.: Efficient Symbolic Signatures for Classifying Craniosynostosis Skull Deformities. In: Workshop of ICCV, pp. 302–313 (2005)
5. Ruiz-Correa, S., Sze, R., Starr, J., Lin, H., Speltz, M., Cunningham, M., Hing, A.: New Scaphocephaly Severity Indices of Sagittal Craniosynostosis: A Comparative Study With Cranial Index Quantifications. Cleft Palate-Craniofacial Journal 43(2), 211–221 (2006)
6. Lee, S.-I., Lee, H., Abbeel, P., Ng, A.: Efficient L_1 Regularized Logistic Regression. In: Proceedings of the 21st National Conference on Artificial Intelligence (2006)
7. Slater, B., Lenton, K., Kwan, M., Gupta, D., Wan, D., Longaker, M.: Cranial sutures: a brief review. Plastic and Reconstructive Surgery 121(4), 170–178 (2008)
8. Gray, H., Carter, H.: Gray's Anatomy. Sterling Publishing (2000)
9. Tibshirani, R., Saunders, M., Rosset, S., Heights, Y., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. R. Statist. Soc. B. 67, 91–108 (2005)
10. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections. Arizona State University (2009), http://www.public.asu.edu/~jye02/Software/SLEP
11. Starr, J., Kapp-Simon, K., Cloonan, Y., Collett, B., Cradock, M., Buono, L., Cunningham, M., Speltz, M.: Pre- and post-surgery neurodevelopment of infants with single-suture craniosynostosis: Comparison with controls. Journal of Neurosurgery (Pediatrics) 107(2), 103–110 (2007)

Retrieval of 4D Dual Energy CT for Pulmonary Embolism Diagnosis

Antonio Foncubierta–Rodríguez[1],
Alejandro Vargas[2], Alexandra Platon[2], Pierre–Alexandre Poletti[2],
Henning Müller[1,2], and Adrien Depeursinge[1,2]

[1] University of Applied Sciences Western Switzerland (HES–SO)
[2] University and University Hospitals of Geneva (HUG), Switzerland
antonio.foncubierta@hevs.ch

Abstract. Pulmonary embolism is a common condition with high short–term morbidity. Pulmonary embolism can be treated successfully but diagnosis remains difficult due to the large variability of symptoms, which are often non–specific including breath shortness, chest pain and cough. Dual energy CT produces 4–dimensional data by acquiring variation of attenuation with respect to spatial coordinates and also with respect to the energy level. This additional information opens the possibility of discriminating tissue with specific material content, such as bone and adjacent contrast. Despite having already been available for clinical use for a while, there are few applications where Dual energy CT is currently showing a clear clinical advantage. In this article we propose to use the additional energy–level data in a 4D dataset to quantify texture changes in lung parenchyma as a way of finding parenchyma perfusion deficits characteristic of pulmonary embolism.

Keywords: 4D texture, pulmonary embolism, dual energy CT.

1 Introduction

Pulmonary embolism (PE) consists of the obstruction of one or several arteries in the lungs as a complication of deep vein thrombosis (DVT). Studies have shown that acute pulmonary embolism mortality rates can reach 75% during initial hospital admission [1] and 30% after 3 years of hospital discharge [2]. Pulmonary embolism is an avoidable cause of death if treated immediately with anticoagulants. Delays in diagnosis of pulmonary embolism have shown to increase the risk of death [3,4], making early diagnosis a key factor for successful treatment.

Schwickert et al. [5] showed that pulmonary embolism induced wedge–shaped pleura–based regions of heterogeneous increased attenuation in unenhanced computed tomography (CT) scans for 54 over 75 patients that were also visible on contrast–enhanced CT. Ganeshan et al. [6] observed that simple 3D texture attributes correlated well with ventilation and vascularization of the lung

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 45–55, 2013.

parenchyma in contrast–enhanced CT scans. Texture information is therefore relevant to quantify pulmonary parenchyma ischemia in CT imaging.

Dual–energy computer tomography (DECT) contains 4D data: three spatial dimensions and the level of x–ray energy between 40 and 140 keV used for image acquisition. Iodine components from the contrast product have differing contrasts at varying energies and are related to the perfusion of the lung parenchyma. Several studies showed the value of DECT to quantify perfusion defects of the lung parenchyma [7,8,9,10,11] using iodine components, which can be derived from CT attenuation at two energy levels of 80 and 140 keV.

This work investigates the use of texture–based image retrieval of 4D DECT, where several energy levels add relevant information about the perfusion of the lung parenchyma. Texture is very hard to visualize for dimensions higher than 2D. Since DECT acquires not only spatially–sampled but also energy–sampled data, visual information is even more difficult to understand in an intuitive way. Computer–assisted analysis of DECT is therefore required. In this article, texture is described by using the wavelet transform together with a visual words approach. The use of texture analysis techniques and challenging 4D data allows to evaluate the possible impact of DECT in early diagnosis of pulmonary embolism.

2 Materials and Methods

This section presents the data of 4D DECT images from PE patients and our approach to 4D texture quantification based on two main ideas: a 3D wavelet transform for multi–scale texture descriptors for each scale and visual words [12,13], as a way of obtaining features based on patterns actually occurring in the data set.

2.1 Dataset

Pulmonary parenchyma ischemia in 4D dual energy CT (DECT) images were identified in collaboration with the emergency radiology of the University Hospitals of Geneva. A small set of 13 currently annotated patients was used to train and test the techniques.

For each patient, the five pulmonary lobes were manually segmented, and the Qanadli index [14] was computed as a measure of the obstruction on a lobe basis. The Qanadli index is computed by adding a score per artery in the lobe: 0 if there is no obstruction, 1 if there is partial obstruction and 2 if the artery is completely obstructed. The maximum value of the Qanadli index varies among lobes, depending on the number of arteries. The Qanadli index value was normalized using the maximum value per lobe, obtaining the percentage of the Qanadli index $Q(\%)$ as a measure of the pulmonary embolism severity.

The images in the dataset contain approximately 300 slices per patient and energy level. Energy levels are sampled from 40 to 140 keV in steps of 10 keV. The total amount of data per patient is approximately $512 \times 512 \times 300 \times 11 = 865.08$ million voxels. Example images can be seen in Figure 1.

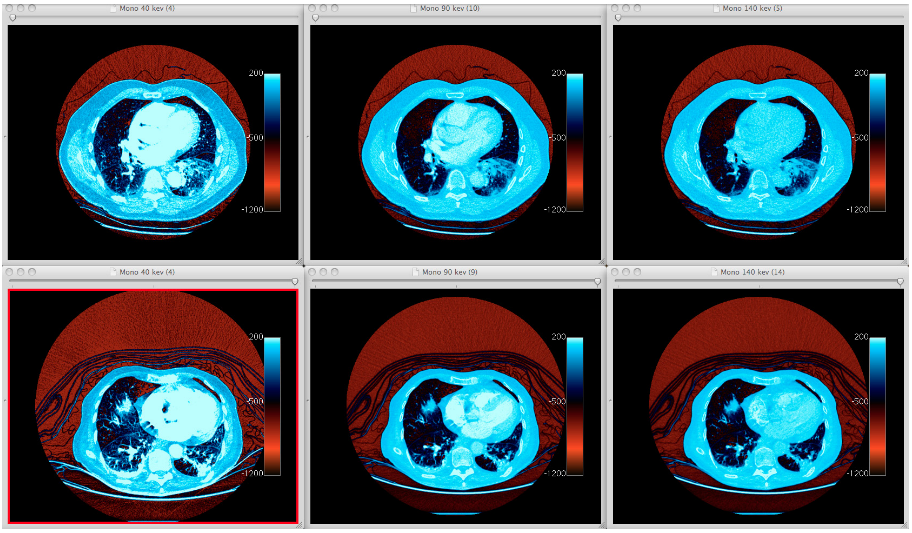


Fig. 1. DECT example with slices at three energy levels per row: 40, 90 and 140 keV

The image resolution is approximately isotropic in the spatial coordinates, with horizontal resolution of $0.83mm/voxel$ and vertical resolution of $1mm/voxel$. This allows for 3D analysis in the spatial domain, whereas the energy domain needs to be taken into account separately, given the different nature of the data. Therefore, experiments were conducted using a spatial, three–dimensional multi–resolution analysis of each energy level. Visual features were then aggregated through energy levels using a visual words approach.

2.2 4D Texture Analysis

Energy of Wavelet Coefficients. The wavelet transform has been widely used to analyze images and videos at multiple resolutions [15,16,17]. In medical imaging, it showed to accurately characterize the lung parenchyma [18,19,12]. The mother wavelet chosen for texture description is the Difference of Gaussians (DoG). Since the image sampling is not fully isotropic, the multi–resolution analysis is based on the Gaussian function g calculated in physical dimensions by scaling the variables x, y and z using the corresponding values for the voxel spacing in each direction (δ_x,δ_y,δ_z):

$$g_{\boldsymbol{\sigma}}(\boldsymbol{x}) = \frac{1}{\sigma_x\sigma_y\sigma_z\sqrt{(2\pi)^3}} e^{-\left(\frac{(x\delta_x)^2}{2\sigma_x^2}+\frac{(y\delta_y)^2}{2\sigma_y^2}+\frac{(z\delta_z)^2}{2\sigma_z^2}\right)}. \tag{1}$$

The choice of using Gaussian functions is based on their good isotropic properties, which allow to analyze image texture without making prior choices of orientation, as opposed to co–ocurrence matrix based methods [20,21].

The extracted coefficients are obtained by using the difference of Gaussians (DoG), which provides a good approximation to the Laplacian of Gaussians or Mexican Hat wavelets when the variance parameters $\sigma_{1,2}$ of two Gaussian functions $g_{1,2}$ satisfy $\sigma_2 \approx 1.6\sigma_1$. Since the functions are calculated in physical dimensions, there is no need for having anisotropic variance parameters of $g_{1,2}$. The resulting wavelets are shown in Equation 2. The number of scales j used for the wavelet transform is five, with j ranging from 1 to 5.

$$\psi_j(\boldsymbol{x}) = g_{\boldsymbol{\sigma_1}}(\boldsymbol{x}) - g_{\boldsymbol{\sigma_2}}(\boldsymbol{x}). \tag{2}$$

$$\boldsymbol{\sigma_2} = 1.6\boldsymbol{\sigma_1}. \tag{3}$$

$$\boldsymbol{\sigma_1} = \boldsymbol{2}^j. \tag{4}$$

The wavelet admissibility condition forces the mother function to have zero–mean. Therefore, the images filtered with $\psi_j(\boldsymbol{x})$ will be bandpass images (i.e., also zero–mean images). For this reason the raw wavelet coefficients are not used, since in the clustering phase all clusters would be located around $\mathbf{0}$. Instead, the energy of the wavelet coefficients averaged in a small neighborhood was chosen as a local feature for describing the images.

Given a voxel identified by its position $\boldsymbol{x}$, and a neighborhood $\mathcal{N}$ with $S_{\mathcal{N}}$ elements, the energy of the wavelet coefficients in the neighborhood is computed as the sum of the squared coefficients within the neighborhood. Since voxels near the boundaries of the image have fewer neighbors than those far from the boundaries, the mean energy over$\mathcal{N}$, $E_w(\boldsymbol{x})$, was chosen instead of the energy, as shown in equation 5. The neighborhood size was kept small (6 × 6 × 6) to maintain a sufficiently local operator on the images. Although the energy was averaged in neighborhoods, the wavelet transform was applied to the complete 3–dimensional image to avoid border effects.

$$E_w(\boldsymbol{x}) = \frac{1}{S_{\mathcal{N}}} \sum_{\boldsymbol{x}\in\mathcal{N}} \psi_j(\boldsymbol{x})^2. \tag{5}$$

Visual Words. The term *texture* often has a fuzzy definition and refers to the (sometimes regular or periodic) visual characteristics of the pixel values within a certain region and their relationships, which is not always explicit to human observers. Since the wavelet transform can describe the transient of the values in the voxel surroundings, a way of aggregating this information for all regions of interest is needed in order to describe the concept of texture which refers to a scale larger than voxels.

Visual words [22,12] have been widely used in image retrieval and image classification for describing image content (or regions of interest). The approach is similar to the bag–of–words approach used for text retrieval or text similarity matching [23].

For each voxel, this technique maps a set of continuous low–level features in a given neighborhood, e.g. gray values or wavelet coefficients, into a compact discrete representation consisting of visual words.

Visual words are cluster centers in the feature space derived from the energy of wavelet coefficients across different scales and energy levels. This guarantees to have a set of visual features actually corresponding to discriminative patterns that do occur in the database. Every image or region is subsequently described by the histogram of the visual words within this region.

Definition 1. *Let $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$ be the set of m descriptors, $f_i \in \mathbb{R}^N$, describing visual characteristics of a given set of images or image regions. A visual vocabulary $W_{\mathcal{F},k} = \{w_1, w_2, \ldots, w_k\}$, with $w_i \in \mathbb{R}^N$ is constructed by grouping the elements of $\mathcal{F}$ into k disjoint subsets or words, and selecting their k centroids w_j with $j \in \{1, \ldots, k\}$.*

The bag–of–visual–words of an image or region I, described by m_I visual descriptors $\{f_1, f_2, \ldots, f_{m_I}\}$, is defined as a vector $h_I = \{c_1, c_2, \ldots, c_k\}$:

$$c_j = \sum_{i=1}^{m_I} g_j(f_i) \quad \forall j \in \{1, \ldots, k\}$$

where

$$g_j(f) = \begin{cases} 1 \text{ if } d(f, w_j) \leq d(f, w_l) & \forall l \in \{1, \ldots, k\} \\ 0 \text{ otherwise} \end{cases}$$

being $d(f, w)$ the distance between two vectors f and w.

2.3 Experimental Setup

First, the images were analyzed using the wavelet transform at two to five scales; the energy of these wavelets in a $6 \times 6 \times 6$ neighborhood was computed. A mask of the lobes was used, storing only the values for the voxels contained in the lung lobes. This process was repeated for all patients and all energy levels.

Given the small number of patients available, the leave–one–patient–out cross–validation method was chosen. Leaving the features from one patient out at a time, k–means clustering was carried out on the rest of the feature space for all the concatenated features to find the visual words; five wavelet scales for each energy level. The number of clusters or visual words ranged from 50 to 150 in steps of 50 as this was computationally feasible.

For all images, the voxels were labeled with the identifier of the nearest cluster center of the vocabulary (e.g., visual word) built from all other patients. For each lobe, the histogram of visual words within the lobe was considered as an instance

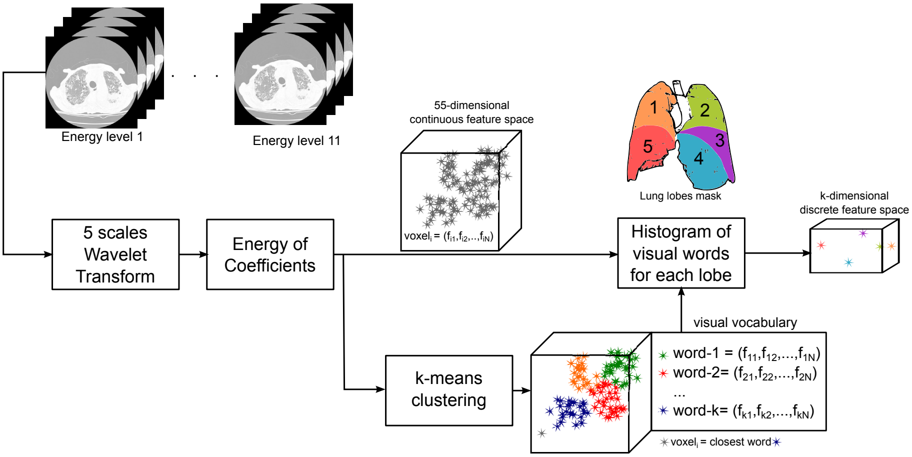

Energy level 1
. . .
Energy level 11
5 scales Wavelet Transform
Energy of Coefficients
55-dimensional continuous feature space
voxel$_i$ = (f_{i1},f_{i2},..,f_{iN})
1
2
3
4
5
Lung lobes mask
Histogram of visual words for each lobe
k-dimensional discrete feature space
visual vocabulary
k-means clustering
word-1 = (f_{11},f_{12},...,f_{1N})
word-2= (f_{21},f_{22},...,f_{2N})
...
word-k= (f_{k1},f_{k2},...,f_{kN})
voxel$_i$ = closest word

Fig. 2. System overview showing the data processing pipeline for each patient

for retrieval using the k nearest neighbors and a Euclidean distance. Figure 2 shows an overview of the complete experimental configuration with a focus on the relative dimensionality of the feature spaces before and after the visual word assignment.

3 Results

In order to evaluate the retrieval precision, a relevance criterion needs to be defined. Since the degree of obstruction of obstruction of the lobe is quantified via the Qanadli index, and given the reduced size of the dataset, it is not possible to define relevance with a strictly equal condition based on the Qanadli index, only.

Table 1. Precision values for a varying number of visual words with 4D data, including all energy levels from 40KeV to 140KeV

Visual Words	Scales	P@1(%)	P@5(%)	P@10(%)
50	1	55	56	56
100	1	58	55	57
150	1	58	56	56
50	2	62	58	55
100	2	62	**62**	**60**
150	2	**63**	**62**	**60**
50	3	58	54	55
100	3	60	59	58
150	3	57	**62**	58
50	5	45	52	51
100	5	57	52	51
150	5	58	52	52

Table 2. Precision values for a varying number of visual words with 3D data only (i.e. 70KeV)

Visual Words	Scales	P@1(%)	P@5(%)	P@10(%)
50	1	60	56	**56**
100	1	57	57	55
150	1	58	56	54
50	2	55	57	55
100	2	58	57	**56**
150	2	**63**	**60**	54
50	3	51	55	53
100	3	49	53	55
150	3	55	55	55
50	5	45	50	52
100	5	51	49	52
150	5	51	52	53

Table 3. Confusion matrices for the best performing configurations for the first retrieved lobes

(a) Dual energy CT (4D) with 2 wavelet scales per energy level and 150 visual words. P@1=63%

	Healthy (%)	Not healthy (%)
Healthy	47.6	52.4
Not healthy	29.5	70.5

(b) Single energy CT (3D) with 2 wavelet scales per energy level and 150 visual words. P@1=63%

	Healthy (%)	Not healthy (%)
Healthy	52.4	47.6
Not healthy	31.8	68.2

In this experiment, all lobes with Qanadli index larger than zero were considered not healthy, whereas only the lobes with Qanadli index equal to zero were considered healthy. Precision at one (P@1), five (P@5) and ten (P@10) were computed for a varying number of visual words, and wavelet scales. Table 1 shows the precision values for the DECT images (4D) whereas Table 1 shows the corresponding results for a single–energy 70KeV CT (3D). This energy level was chosen because it is the standard energy level used by radiologists to diagnose pulmonary embolism.

Table 3 shows the confusion matrices for the best performing configurations for the top retrieved lobes. Tables 4 and 5 show the corresponding matrices for precision of the first five and the first ten retrieved lobes.

Table 4. Confusion matrices for the best performing configurations for the first five retrieved lobes

(a) Dual energy CT (4D) with 2 scales per energy level and 100 visual words. P@5=62%

	Healthy (%)	Not healthy (%)
Healthy	43.8	56.2
Not healthy	29.5	70.5

(b) Dual energy CT (4D) with 2 wavelet scales per energy level and 150 visual words. P@5=62%

	Healthy (%)	Not healthy (%)
Healthy	43.8	56.2
Not healthy	29.9	70.1

(c) Dual energy CT (4D) with 3 wavelet scales per energy level and 150 visual words. P@5=62%

	Healthy (%)	Not healthy (%)
Healthy	44.8	55.2
Not healthy	29.1	70.9

Table 5. Confusion matrices for the best performing configurations for the first ten retrieved lobes

(a) Dual energy CT (4D) with 2 wavelet scales per energy level and 100 visual words. P@10=60%

	Healthy (%)	Not healthy (%)
Healthy	42.4	57.6
Not healthy	31.1	68.9

(b) Dual energy CT (4D) with 2 wavelet scales per energy level and 150 visual words. P@10=60%

	Healthy (%)	Not healthy (%)
Healthy	40	60
Not healthy	31.1	68.9

4 Discussion and Conclusions

This article presents an approach for solid texture analysis on 4D dual energy CT data to better detect and quantify pulmonary embolisms for a more efficient and faster treatment in emergency radiology by retrieving visually similar cases to support clinical decisions. To the best of the author's knowledge such an apprach has not yet been described in the literature.

Results in Section 3 show that four dimensional texture contains patterns related to the pulmonary embolism severity. Six out of the seven best performing configurations are obtained with dual energy data. The number of visual words and wavelet scales also have an impact on the accuracy. The optimal configuration for the multiscale framework was found for 2 scales, and the number of visual words that performed best was 150 visual words, which agrees with [12].

The confusion matrices from Tables 3, 4 and 5 show that the accuracy for PE lobes is much better than for healthy lobes. There are two possible explanations for these results. First, the dataset contains only patients with PE, which can affect the lung parenchyma of the lobes without embolisms which can be overloaded by redirected blood flows. Second, healthy lobes can contain a very rich set of patterns possibly linked to other diseases.

In this paper, a novel use of dual energy CT data is proposed. The limitations of visual inspection for higher dimensional data are underlined and an automatic analysis is suggested as an aid for clinicians for detection and quantization of pulmonary embolisms. Despite the small size of the dataset, the relationship between four dimensional texture and pulmonary embolism severity was shown, proposing a new diagnosis tool that may help particularly emergency radiologists to diagnose and quantify pulmonary embolism and potentially reduce mortality through quicker and more accurate treatment.

A small scale evaluation was performed as there are currently no large data sets with corresponding ground truth available and we only started acquiring new cases with a DECT protocol. This only allows a few basic conclusions on the techniques employed. However, the results state a baseline for future work and point at the use of medical image computing as a means to overcome the limitations of human perception and understanding of high dimensional image data. Nevertheless, computational complexity is very high: the offline feature extraction took approximately 4 hours per patient on a 24–core machine with 96 gigabytes of main memory.

We currently work on creating a larger database to allow for a better evaluation and also for more training data for our system. The future dataset is foreseen to contain control cases that will allow for a better characterization of the healthy lobes.

Acknowledgments. This work was partially supported by the Swiss National Science Foundation (FNS) in the MANY project (grant 205321–130046), the EU 7^{th} Framework Program in the context of the Khresmoi project (FP7–257528), and the Center for Biomedical Imaging (CIBM).

References

1. Goldhaber, S.Z., Visani, L., Rosa, M.D.: Acute pulmonary embolism: clinical outcomes in the international cooperative pulmonary embolism registry (icoper). The Lancet 353(9162), 1386–1389 (1999)
2. Anderson, F.A.J., Wheeler, H.B., Goldberg, R.J., Hosmer, D.W., Patwardhan, N.A., Jovanovic, B., Forcier, A., Dalen, J.E.: A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism: The worcester dvt study. Archives of Internal Medicine 151(5), 933–938 (1991)
3. Alonso-Martínez, J., Sánchez, F.A., Echezarreta, M.U.: Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism. European Journal of Internal Medicine 21(4), 278–282 (2010)
4. Ozsu, S., Oztuna, F., Bulbul, Y., Topbas, M., Ozlu, T., Kosucu, P., Ozsu, A.: The role of risk factors in delayed diagnosis of pulmonary embolism. The American Journal of Emergency Medicine 29(1), 26–32 (2011)
5. Schwickert, H.C., Schweden, F., Schild, H.H., Piepenburg, R., Düber, C., Kauczor, H.U., Renner, C., Iversen, S., Thelen, M.: Pulmonary arteries and lung parenchyma in chronic pulmonary embolism: preoperative and postoperative CT findings. Radiology 191(2), 351–357 (1994)
6. Ganeshan, B., Miles, K.A., Young, R.C.D., Chatwin, C.R.: Three–dimensional selective–scale texture analysis of computed tomography pulmonary angiograms. Investigative Radiology 43(6), 382–394 (2008)
7. Chae, E.J., Seo, J.B., Jang, Y.M., Krauß, B., Lee, C.W., Lee, H.J., Song, K.S.: Dual–energy CT for assessment of the severity of acute pulmonary embolism: Pulmonary perfusion defect score compared with CT angiographic obstruction score and right ventricular/left ventricular diameter ratio. American Journal of Roentgenology 194(3), 604–610 (2010)
8. Lee, C., Seo, J., Song, J.W., Kim, M.Y., Lee, H., Park, Y., Chae, E., Jang, Y., Kim, N., Krauß, B.: Evaluation of computer–aided detection and dual energy software in detection of peripheral pulmonary embolism on dual-energy pulmonary CT angiography. European Radiology 21(1), 54–62 (2011)
9. Nakazawa, T., Watanabe, Y., Hori, Y., Kiso, K., Higashi, M., Itoh, T., Naito, H.: Lung perfused blood volume images with dual–energy computed tomography for chronic thromboembolic pulmonary hypertension: Correlation to scintigraphy with single–photon emission computed tomography. Journal of Computer Assisted Tomography 35(5), 590–595 (2011)
10. Thieme, S.F., Becker, C.R., Hacker, M., Nikolaou, K., Reiser, M.F., Johnson, T.R.C.: Dual energy CT for the assessment of lung perfusion—correlation to scintigraphy. European Journal of Radiology 68(3), 369–374 (2008)
11. Thieme, S.F., Johnson, T.R.C., Lee, C., McWilliams, J., Becker, C.R., Reiser, M.F., Nikolaou, K.: Dual–energy CT for the assessment of contrast material distribution in the pulmonary parenchyma. American Journal of Roentgenology 193(1), 144–149 (2009)
12. Foncubierta-Rodríguez, A., Depeursinge, A., Müller, H.: Using Multiscale Visual Words for Lung Texture Classification and Retrieval. In: Müller, H., Greenspan, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2011. LNCS, vol. 7075, pp. 69–79. Springer, Heidelberg (2012)
13. Avni, U., Greenspan, H., Konen, E., Sharon, M., Goldberger, J.: X–ray categorization and retrieval on the organ and pathology level, using patch–based visual words. IEEE Transactions on Medical Imaging 30(3), 733–746 (2011)

14. Qanadli, S.D., El Hajjam, M., Vieillard-Baron, A., Joseph, T., Mesurolle, B., Oliva, V.L., Barré, O., Bruckert, F., Dubourg, O., Lacombe, P.: New CT index to quantify arterial obstruction in pulmonary embolism. American Journal of Roentgenology 176(6), 1415–1420 (2001)
15. Nevel, A.V.: Texture classification using wavelet frame decompositions. In: Conference Record of the Thirty–First Asilomar Conference on Signals, Systems & Computers, vol. 1, pp. 311–314 (November 1997)
16. Smith, J.R., Lin, C.Y., Naphade, M.: Video texture indexing using spatio–temporal wavelets. In: Proceedings of the International Conference on Image Processing, vol. 2, pp. 437–440 (September 2002)
17. Chenouard, N., Unser, M.: 3D steerable wavelets and monogenic analysis for bioimaging. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 2132–2135 (April 2011)
18. Korfiatis, P., Skiadopoulos, S., Sakellaropoulos, P., Kalogeropoulou, C., Costaridou, L.: Automated 3D Segmentation of Lung Fields in Thin Slice CT Exploiting Wavelet Preprocessing. In: Kropatsch, W.G., Kampel, M., Hanbury, A. (eds.) CAIP 2007. LNCS, vol. 4673, pp. 237–244. Springer, Heidelberg (2007)
19. Korfiatis, P., Kalogeropoulou, C., Karahaliou, A., Kazantzi, A., Skiadopoulos, S., Costaridou, L.: Texture classification–based segmentation of lung affected by interstitial pneumonia in high–resolution CT. Medical Physics 35(12), 5290–5302 (2008)
20. Korfiatis, P.D., Kalogeropoulou, C., Karahaliou, A.N., Kazantzi, A.D., Costaridou, L.I.: Vessel tree segmentation in presence of interstitial lung disease in MDCT. IEEE Transactions on Information Technology in Biomedicine 15(2), 214–220 (2011)
21. Kovalev, V.A., Kruggel, F.: Texture anisotropy of the brain's white matter as revealed by anatomical MRI. IEEE Transactions on Medical Imaging 26(5), 678–685 (2007)
22. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV 2003, vol. 2, pp. 1470–1477. IEEE Computer Society, Washington, DC (2003)
23. van Rijsbergen, C.J.: Information Retrieval. Prentice Hall, Englewood Cliffs (1979)

Immediate ROI Search for 3-D Medical Images

Karen Simonyan[1], Marc Modat[2], Sebastien Ourselin[2], David Cash[2],
Antonio Criminisi[3], and Andrew Zisserman[1]

[1] University of Oxford, UK
{karen,az}@robots.ox.ac.uk
[2] Centre for Medical Image Computing, University College London, UK
{m.modat,s.ourselin,d.cash}@ucl.ac.uk
[3] Microsoft Research, Cambridge, UK
antcrim@microsoft.com

Abstract. The objective of this work is a scalable, real-time, visual search engine for 3-D medical images, where a user is able to select a query Region Of Interest (ROI) and automatically detect the corresponding regions within all returned images.

We make three contributions: (i) we show that with appropriate off-line processing, images can be retrieved and ROIs registered in real time; (ii) we propose and evaluate a number of scalable exemplar-based image registration schemes; (iii) we propose a discriminative method for learning to rank the returned images based on the content of the ROI. The retrieval system is demonstrated on MRI data from the ADNI dataset, and it is shown that the learnt ranking function outperforms the baseline.

Keywords: Immediate structured search, visual search, ROI, exemplar-based registration, learning to rank.

1 Introduction

Throughout the last decade there has been a rapid growth of medical image repositories. Medical images and corresponding clinical cases, stored in these large collections, capture a wide range of disease population variability due to numerous covariates (diagnosis, age, co-morbidities, etc). Instant image retrieval from such repositories could be of great value for clinical practice, e.g. by providing a "second opinion" based on the corresponding diagnostic information or course of treatment. Apart from the processing speed, another important aspect of a practical retrieval system is the ability to focus the search on a particular part (structure) of the image which is of most interest.

This paper addresses the problem of immediate structured image retrieval in large repositories of 3-D medical images. Given a query 3-D image (e.g. from a new patient we wish to diagnose) and a user-drawn Region Of Interest (ROI) in it, we seek to retrieve repository images with the ROI automatically located, and rank them based on a clinically relevant score, driven by the content of the ROI. Instant ROI localisation in large repositories is achieved by off-line preprocessing of the repository based on fast image registration. Figure 1 shows screenshots of our brain MRI retrieval system.

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 56–67, 2013.

The contribution of this paper is three-fold. First, we show that 90 (and potentially more) images can be retrieved and ROIs registered in real time. Second, we present and evaluate several modifications of scalable exemplar-based image registration. Finally, we propose a technique for learning to rank the retrieved ROI. We envisage a number of applications of the proposed framework, and discuss three of them below.

Atrophy-Aware Brain MRI Retrieval. Structural MRI data has been shown to provide reliable quantification of the atrophy process in the brain caused by Alzheimer's disease (AD) [5] or other neurodegenerative disorders. There are numerous natural history studies, the Alzheimer's Disease Neuroimaging Initiative (ADNI) [9] being the most prominent. The immediate visual search engine can aid in differential diagnosis, as there are discriminating patterns between numerous forms of dementia. The ability to focus the search on specific anatomical regions, that have been identified as being sensitive to the disease, is an important advantage of structured (ROI-level) visual search as opposed to retrieval based on global cues. For example, the hippocampal deterioration is increasingly being considered as a way of identifying subjects who have a higher risk of developing AD. Providing the images with relevant ROI and their respective diagnosis to clinicians will aid in their decision process. We give an implementation of a search-engine for this application in this paper.

Lesion Retrieval in CT Scans. The wide application of computed tomography to lesion detection (e.g. liver or kidney lesions) has led to the collection of large quantities of imaging data together with corresponding clinical reports. Recently, image retrieval frameworks [11] have been proposed, which can help clinicians to search for similar lesions in image repositories. However, such methods do not take into account the relative location of the lesion inside the liver, which can be an important search criterion (e.g. the query "find all visually similar lesions in the same part of liver"). To process such queries, the geometrical correspondence of a query ROI in target images should be quickly obtained, which can be done using exemplar-based registration employed in the proposed framework.

Image Quality Control. Another application of the visual search engine is the quality control of incoming images for research studies and clinical trials. This predominantly manual task is one of the most time consuming areas of the processing pipeline. Even though the failure rate is low, all data needs to be reviewed by radiologists and images with poor quality must be excluded from the analysis. Typically, this task consists of a careful qualitative review of each image independently. If the reviewers were provided with a visual search engine to retrieve similar (on ROI-level) images from the repository and the outcomes of their image quality review, this would speed up the process dramatically.

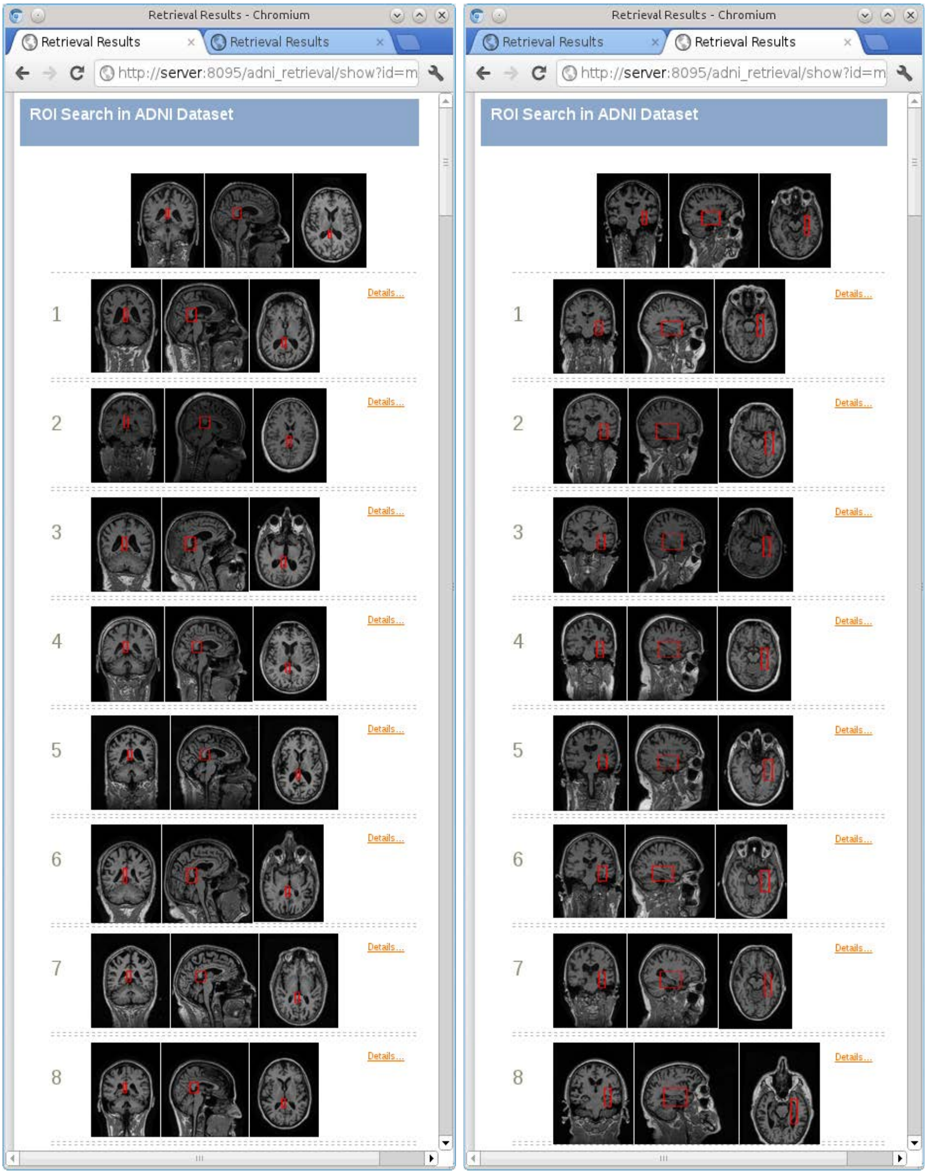


Fig. 1. Screenshots of our brain MRI retrieval system for two different queries and the top eight retrieval results. The system is accessed via a conventional Web browser. The top of the page shows the three orthogonal views of the user-specified query volume and axis-aligned ROI outlined in red. The high ranked retrieved volumes with the corresponding ROI (in red) are shown below.

1.1 Related Work

The problem of content-based medical image retrieval has a vast literature. Most conventional approaches [10] consist in retrieving images that are *globally* similar to the query image. Recently, the problem of ROI-level search has been addressed in [1,2,14]. In [1,2], an approach related to that of [15] was employed, which consists in computing an ROI descriptor vector (known as a bag of visual words), and retrieving images based on the distance between query and target ROI descriptors. It should be noted that such a technique discards information about the spatial location of the query ROI in an image. While this can be an advantage when searching in natural images or videos [15], in the case of medical images this can result in the retrieved ROIs lying in completely different anatomical locations, which is often undesirable. To circumvent this, augmenting of visual desriptors with their spatial location was proposed in [1]. A more principled approach is taken in the framework of [14], which makes use of the fact that medical images are usually acquired under standardised protocols with a fixed viewpoint, field of view, etc. It allows one to quickly compute registrations between query and repository images, making the target ROI detection trivial. We build upon their framework and review it in more detail below.

Outline. The rest of the paper is organised as follows. Section 2.1 contains the details of our brain MRI dataset and the retrieval system implementation. In Sect. 3 we present several exemplar-based registration techniques and evaluate them. In Sect. 4 we describe a learning to rank framework.

2 ROI Retrieval Framework

To enable immediate ROI retrieval at run time, processing is divided into off-line and on-line parts, as summarised in Fig. 2. A similar approach was previously applied to 2-D X-ray images retrieval [14]. The key idea is to pre-compute registrations between repository images off-line, so that at run time the correspondences of a query ROI in target images can be obtained immediately if the query image is taken from the repository. Once the regions of interest have been aligned in repository images, they can be ranked based on an application-specific clinically relevant score.

If a query image is not in the repository, it is added there by registering it with repository images. This brings up the issue of computational efficiency in the case of large datasets. In [14] an efficient exemplar-based registration technique was proposed to solve this problem. It requires only a small fixed number of registrations to exemplar images to be computed. The transformations to the rest of the repository is then obtained by computationally cheap transform composition. It should be noted that registration with exemplar images can be performed using *any* off-the-shelf method suitable for a particular type of images. In this paper, we apply the framework to a large dataset of brain MRI scans, where images exhibit similar fields of view.

1. **On-line (given a user-specified query volume and ROI bounding box)**
   - Using the pre-computed registration and transform composition (Sect. 3), compute the ROIs corresponding to the query ROI in all repository images.
   - Rank the retrieved ROIs using a clinically meaningful ranking function of choice (Sect. 4).
2. **Off-line (pre-processing)**
   - Compute registration between exemplar images and all other images (Sect. 3)

Fig. 2. The retrieval algorithm outline

2.1 Dataset and Implementation Details

Our dataset consists of 90 brain MRI scans randomly selected from the ADNI dataset [9]. The dataset contains an equal number of images (30) of each of the three subject groups: Alzheimer's disease, control, and MCI (mild cognitive impairment). For the evaluation of the methods proposed in the paper, for each of these images we computed the "gold standard" parcellation and pairwise registrations. We note that this is not required for the functioning of the proposed search engine. The parcellation into 83 brain anatomical structures was performed using the method of [4]. The non-rigid registration has been performed using the Free-Form Deformation approach [12]. Briefly, it consists of a cubic B-Spline parametrisation model where the Normalised Mutual Information (NMI) is used as a measure of similarity. We used an efficient implementation [8] that is freely available as a part of the NiftyReg package. We provide a default ranking function based on the contents of the ROI (the ranking function can also be learnt, Sect. 4). For the default, we measure the χ^2 distance between the brain tissue type distributions in query and target ROI. The distributions were computed using the GMM-based probabilistic segmentation algorithm [3].

Our retrieval system is implemented as a Web-based application, which can be accessed from any device equipped with a Web browser (thin client paradigm). The system is split into a front-end and back-end. The front-end, implemented in Python and JavaScript, allows a user to select a query volume, specify arbitrary axis-aligned ROI in it, and explore the retrieval results. The back-end is currently implemented in unoptimised Python, leaving a lot of room for potential speed-up. The average ROI registration time using five exemplar images (Sect. 3) on a single CPU core is 0.06s per image, which allows for retrieval of hundreds of MRI volumes under 1s when rolled out on a multi-core server.

In certain use cases, using multiple query ROI can be beneficial, as it would allow one to select several relevant areas in a query image. Here we consider a single query ROI, the extension to multiple ROI being rather straightforward. We also restrict the ROI to be an axis-aligned 3-D bounding box, but in general any ROI shape is possible.

3 Exemplar-Based Registration

Carrying out non-rigid registration of the query image with each of the target images scales badly with the number of repository images as 3-D image registration is computationally complex, and the number of registrations equals the number of images. Moreover, storing all pairwise registrations is prohibitive due to high storage requirements of non-rigid transforms (e.g. B-spline warps computed over a dense 3-D grid).

The key idea behind scalable exemplar-based registration is that instead of registering a query image with each of the repository images by pairwise registration, the query is registered with only a few *fixed* images (called exemplars), which effectively define several reference spaces. The remaining repository images are already pre-registered with exemplars, so they can be registered with the query by composing the two transforms. Finally, to obtain a single correspondence from several exemplars, the composed transforms are aggregated. The exemplar-based registration is schematically illustrated in Fig. 3 (left).

More formally, for a dataset of N images, a query image I_q is registered with only a subset of $K = \text{const}$ exemplar images, which results in K transforms $T_{q,k},\ k = 1 \ldots K$. The transformations $T_{k,t}$ between an exemplar I_k and each of the remaining repository images I_t are pre-computed. Then the transformation between images I_q and I_t can be obtained by composition of transforms (computed using different exemplars) followed by aggregation:

$$T_{q,t}(\mathbf{x}) = f\left(\{T_{k,t} \circ T_{q,k}\}\right)(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is a point in the query image and f is the aggregation function.

The advantage of exemplar-based registration scheme is that for a query image only $K \ll N$ registrations should be computed, and the transform composition complexity is negligible. Thus pairwise registrations between all images can be computed in $O(KN)$ rather than $O(N^2)$. The same estimates apply to the storage requirements for the computed registrations, which allows them to be stored in RAM for fast access. Compared to group-wise registration algorithms, transform composition does not rely on the computation of a group mean model, and is scalable in the case of rapidly growing datasets. Additionally, the use of several transformations instead of one improves the registration robustness.

There are two important choices to make: how to select the exemplars and how to define the function f, aggregating the transforms obtained using different exemplars. In [14] the exemplars were selected randomly, and the aggregation was performed by taking a median. In both cases the accuracy of registration is not taken into account, which can potentially lead to the selection of exemplars which can not be accurately registered with other images. If the ratio of such exemplars is large, the median filter will not be able to recover the correspondence.

Here we investigate different ways of exemplar selection and transform aggregation. We assume that the registration error $d_{ij} = d(I_i, I_j)$ for two images I_i and I_j belongs to the range $[0, 1]$ with 0 corresponding to a perfect registration.

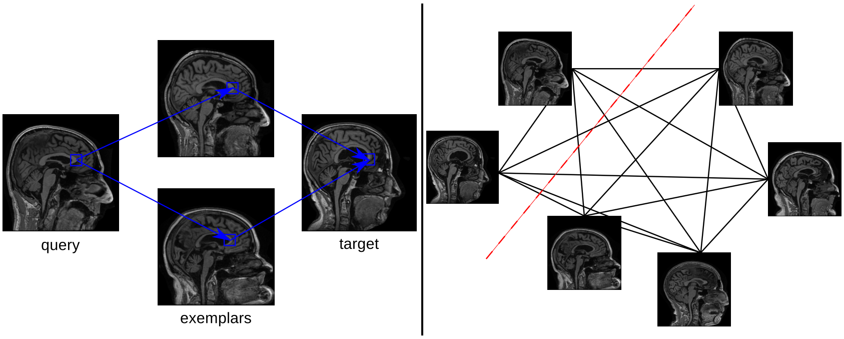


Fig. 3. *Left*: Exemplar-based registration. *Right*: Repository graph. The red dashed line illustrates the graph cut into exemplars and other images.

In general, the error can be computed using different cues (intensity, deformation field smoothness, re-projection error, etc.). In our experiments, we employed inverse normalised mutual information, rescaled to the $[0, 1]$ range.

3.1 Exemplar Images Selection

The objective of exemplar selection is to pick a fixed number (K) of repository images, such that they can be accurately registered with the remaining ones. In this section we describe several ways of deterministic (non-random) exemplar selection from a set of images based on their pairwise registration errors. For instance, the exemplars selection can be carried out on the full repository or its subset. We stress that deterministic exemplar selection (including pairwise registrations) is performed off-line and has no impact on the query times; random exemplar selection does not require any additional processing at all.

It is natural to represent an image repository as a fully-connected graph with vertices corresponding to images and edges weighted by the registration errors, as shown in Fig. 3 (right). We employ the repository graph formalism to describe the objective functions for deterministic exemplar selection.

Min-Sum Selection. The task of exemplar images selection can be formulated as a min-cut problem on a repository graph. Indeed, we aim at splitting the set of vertices (images) into two partitions such that the sum of edge weights (registration errors) between vertices lying in these partitions is minimal. The resulting optimisation problem is as follows:

$$\alpha = \arg\min_{\alpha} \sum_{i,j} \alpha_i (1 - \alpha_j) d_{ij}, \quad \text{s.t.} \ \sum_i \alpha_i = K \tag{2}$$

where $\alpha \in \{0, 1\}^N$ is a binary vector such that $a_k = 1$ iff k-th repository image is selected as an exemplar. The optimisation of (2) is NP-hard. While efficient approximate solutions exist [7], we leave their evaluation for future work. In this paper, we evaluate exemplar selection based on the following simplified objective:

$$\alpha = \arg\min_{\alpha} \sum_{i,j} \alpha_i d_{ij}, \quad \text{s.t.} \sum_i \alpha_i = K \tag{3}$$

which corresponds to selecting K exemplars such that the sum of edge weights between them and *all* other images (including other exemplars) is minimal. The solution of (3) can be easily obtained by ranking the images in the ascending order of $q_i = \sum_j d_{ij}$ and then selecting the top-K images as exemplars.

Spectral Clustering Selection. Another approach to deterministic exemplar selection is based on clustering the repository images into K clusters followed by the selection of a single exemplar in each of these clusters. Given the pairwise similarity matrix $s_{ij} = 1 - d_{ij}$, we use the normalised cuts technique [13] to split the vertices (images) into a set of classes such that the similarity between images in different clusters is small, and the similarity between images in the same cluster is large. This corresponds to computing a normalised cut in the repository graph. Once the images are divided into clusters, a single exemplar is selected in each of the clusters as the image with minimal sum of registration errors to the others (3).

3.2 Shortest Path Aggregation

Once the exemplars are selected and fixed, the way of aggregating several registrations into one should be defined (function f in (1)). In general, taking the mean or median does not account for the exemplars registration error, which can be large for certain pairs of query and target images. One of the possible ways to account for these errors is to pick a single registration which corresponds to the shortest path in the graph from the query to the target vertices and goes through exactly one exemplar (Fig. 3, left). In other words, for a given (query, target) pair of images, only one exemplar is selected, which has the lowest sum of registration errors with these images:

$$\begin{aligned} f(q,t)(\mathbf{x}) &= \left(T_{s,t} \circ T_{q,s}\right)(\mathbf{x}), \\ s &= \arg\min_k d_{qk} + d_{kt} \end{aligned} \tag{4}$$

3.3 Evaluation

In this section, we compare the registration accuracy of different combinations of exemplar selection and transform aggregation techniques. For exemplar selection, we consider random selection ("rand"), "min-sum" selection, and spectral clustering selection (Sect. 3.1). For transform aggregation, "median", "mean", and the shortest path exemplar ("single", Sect. 3.2) are compared. The evaluation was performed on the dataset described in Sect. 2.1, which was randomly split into 45 training and 45 testing images. Exemplar selection was performed on the training set, registration evaluation – on the test set. The experiment was repeated three times. For each pair of test images, the accuracy of registration was assessed using two criteria. First, we measured the mean distance (in mm) between points projected using pairwise (between query and target) and exemplar-based transformations. The measure describes how different exemplar-based registration is from the pairwise registration. The points were selected to be the centers of mass of the 83 parcellated anatomical structures. The second measure is the mean overlap ratio (Jaccard coefficient) of 83 anatomical structure bounding boxes, projected from the query image to the target image, with the bounding boxes in the target image. We used the bounding boxes of the parcellated anatomical structure volumes instead of the volumes themselves because it more closely follows the search engine use case scenario, where we operate on the level of bounding boxes. We note that this measure is noisy due to the possible inaccuracies of the "gold standard" parcellation. In Table 1 we report the mean and standard deviation of the two measures across all test image pairs for different number K of exemplar images.

Based on the presented results, we can conclude that all three exemplar selection methods (including the random choice) exhibit similar levels of performance when coupled with robust median aggregation. Aggregation based on shortest path selection performs worse, and the mean aggregation is the worst. The reason for such a behaviour could be that the global registration error, which we used for exemplar selection, does not account for the local inaccuracies. Another reason for similar performance can be the lack of strong image variation in our dataset. At the same time, using a single exemplar ($K = 1$) results in worse accuracy compared to several exemplar images. The accuracy of exemplar-based registration with median aggregation is at the same level as that of pairwise registration without exemplars. The average distance between the points projected using the two registrations is less than 1.4 mm. Considering its low computational complexity, in our practical implementation we used the randomised selection of $K = 5$ exemplars and the median aggregation of the composed transforms.

Table 1. Exemplar-based registration accuracy. The overlap ratio of pairwise registration (without exemplars) is 0.568 ± 0.076. For the overlap ratio higher is better, and for the distance smaller means closer to the direct registration without exemplars.

exemplar	aggregation	overlap ratio			distance (mm)		
		$K = 1$	$K = 5$	$K = 7$	$K = 1$	$K = 5$	$K = 7$
rand	mean	0.555 ±0.072	0.532 ± 0.073	0.53 ± 0.073	2.04 ±0.28	1.44 ± 0.22	1.38 ± 0.21
	median		0.569 ± 0.076	**0.571 ± 0.076**		1.45 ± 0.23	1.37 ± 0.23
	single		0.557 ± 0.073	0.559 ± 0.073		1.99 ± 0.26	1.98 ± 0.25
min-sum	mean	0.557 ±0.072	0.531 ± 0.072	0.529 ± 0.072	1.94 ±0.26	1.42 ± 0.22	1.37 ± 0.22
	median		0.569 ± 0.076	0.57 ± 0.076		1.43 ± 0.23	1.36 ± 0.23
	single		0.558 ± 0.072	0.556 ± 0.072		1.94 ± 0.26	2.00 ± 0.32
cluster	mean		0.531 ± 0.072	0.529 ± 0.072		1.44 ± 0.22	1.39 ± 0.22
	median		0.569 ± 0.076	0.57 ± 0.076		1.45 ± 0.23	1.38 ± 0.23
	single		0.556 ± 0.072	0.556 ± 0.072		2.03 ± 0.32	2.03 ± 0.31

4 Learning to Rank Retrieved ROI

In this section we propose a framework for discriminative learning of ROI ranking. In general, we aim at automatically reproducing the ROI ranking provided by a clinician in the form of preference constraints. That is, for a particular query image and ROI the expert selects *pairs* of retrieved images (with corresponding ROI) such that the first image of the pair should be ranked higher than the second one. More formally, the annotation is represented as a set of triplets: $\{(I_q, R_q), (I_h, R_h), (I_l, R_l)\}_i$ where (I_q, R_q) are the query image and ROI, and the (image, ROI) pair (I_h, R_h) should be ranked higher than (I_l, R_l).

We propose to learn a distance in the space of ROI such that ranking of the retrieved ROI based on their distance to the query ROI satisfies the ground-truth preference constraints with a margin [6], i.e.

$$d(R_q^i, R_h^i) + 1 < d(R_q^i, R_l^i) \tag{5}$$

where d is the distance between ROIs in the feature space (we omit image notation for brevity). We constrain the distance to be a generalised squared Mahalanobis distance of the form:

$$d_A(R_u, R_v) = (\phi_u - \phi_v)^T A(\phi_u - \phi_v), \tag{6}$$

where $A \succeq 0$ is a positive semi-definite matrix to be learnt, and ϕ_u is the feature vector of the ROI R_u (discussed later). It can be shown that the distance (6) equals the squared Euclidean distance in the projected space defined by a projection matrix W such that $W^T W = A$. Taking into account the preference constraints (5), the large-margin learning objective takes the form:

$$A = \arg\min_{A \succeq 0} \sum_i \max\left(d_A(R_q^i, R_h^i) - d_A(R_q^i, R_l^i) + 1, 0\right) + \frac{\lambda}{2}\|A\|_F^2 \tag{7}$$

where the first term is a sum of ranking hinge losses, and the second term is a Frobenius (Euclidean) norm of the matrix A. The parameter $\lambda > 0$, balancing the two parts, is selected on a validation set. The max-margin formulation is closely related to the LMNN formulation of [17]. The cost function (7) is strongly convex and can be efficiently optimised by projected stochastic sub-gradient method.

ROI Feature Vector. The proposed ROI distance learning framework is generic and can be applied to different ROI representations. Here we consider the "bag of words" representation [15] which consists in computing visual descriptors inside the ROI, assigning them to a nearest cluster ("visual word" from a vocabulary), and then accumulating the assignments inside an ROI into a histogram of words. The visual words vocabulary is computed using k-means clustering. The visual descriptors, the corresponding vocabulary, and visual word assignments can be pre-computed off-line. At query time, only the histogram over ROI should be computed, which can be done quickly using integral volumes.

Evaluation. Here we describe a preliminary experiment which was carried out on the dataset of Sect. 2.1. Considering that expert-annotated preference constraints are not currently available for our data (it is a subject of future research), we used the clinical diagnosis class labels ("AD"(Alzheimer's) and "Control") to generate the constraints of the form (5). Namely, we constrain the distance between same-class ROIs to be smaller than the distance between different-class ROI. As the ROI, we used a bounding box in the hippocampal area, which is known to be relevant to the Alzheimer's disease. Bag of words was computed using single-scale dense textons [16] of size $3 \times 3 \times 3$mm which were quantised into 512 visual words, leading to a 512-D ROI feature vector. We randomly selected 30 images for training, 10 for validation, and 20 for testing. The experiment was repeated three times. Mean average precision of retrieval was measured to be 58.8% using Euclidean distance between feature vectors (i.e. $A = I$), and 63.8% using the learnt distance. This shows that metric learning can indeed improve the retrieval performance. We believe that with appropriate preference constraints annotation and more sophisticated visual features, the results of the proposed learning-to-rank framework can be further improved.

5 Summary

In this paper we presented a practical structured image search framework, capable of instant retrieval of brain MRI volumes and corresponding regions of interest from large datasets. Fast ROI alignment in repository images was made possible using scalable exemplar-based registration technique.

The evaluation of different exemplar-based registration methods has shown that random exemplar image selection coupled with robust median transform aggregation achieves registration accuracy on par with optimised exemplar selection and pairwise registration without exemplars. We note that in the case of diverse non-uniform image datasets, deviant images can be unrepresented in the exemplar set. In that case, our conclusions might not be immediately applicable.

Finally, we have presented a discriminative distance learning framework for ranking retrieved ROIs. It was demonstrated that it can indeed improve the ranking performance. It should be noted that while the proposed ranking function has been learnt on a specific anatomical area (hippocampal area), the same approach can be used to learn more generic ranking functions.

A web-based demo of 3-D ROI retrieval framework will be made available at http://www.robots.ox.ac.uk/~vgg/research/med_search/

Acknowledgements. This work was supported by Microsoft Research PhD Scholarship Program, ERC grant VisRec no. 228180, and CBRC grant no. 168.

References

1. Avni, U., Greenspan, H., Konen, E., Sharon, M., Goldberger, J.: X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. IEEE Trans. Med. Imag. 30(3), 733–746 (2011)

2. Burner, A., Donner, R., Mayerhoefer, M., Holzer, M., Kainberger, F., Langs, G.: Texture Bags: Anomaly Retrieval in Medical Images Based on Local 3D-Texture Similarity. In: Müller, H., Greenspan, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2011. LNCS, vol. 7075, pp. 116–127. Springer, Heidelberg (2012)
3. Cardoso, M.J., Clarkson, M.J., Ridgway, G.R., Modat, M., Fox, N.C., Ourselin, S.: LoAd: A locally adaptive cortical segmentation algorithm. NeuroImage 56(3), 1386–1397 (2011)
4. Cardoso, M., Modat, M., Ourselin, S., Keihaninejad, S., Cash, D.: Multi-STEPS: Multi-label similarity and truth estimation for propagated segmentations. In: IEEE Workshop on Math. Meth. in Biomed. Im. Anal., pp. 153–158 (2012)
5. Jack, C.R., Shiung, M.M., Gunter, J.L., O'Brien, P.C., Weigand, S.D., Knopman, D.S., Boeve, B.F., Ivnik, R.J., Smith, G.E., Cha, R.H., Tangalos, E.G., Petersen, R.C.: Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 62(4), 591–600 (2004)
6. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM SIGKDD Int. Conf. on Knowl. Disc. and Data Mining, pp. 133–142. ACM Press, New York (2002)
7. Lim, Y., Jung, K., Kohli, P.: Energy Minimization under Constraints on Label Counts. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 535–551. Springer, Heidelberg (2010)
8. Modat, M., Taylor, Z., Barnes, J., Hawkes, D., Fox, N., Ourselin, S.: Fast free-form deformation using graphics processing units. Comp. Meth. and Prog. Biomed. 98(3), 278–284 (2010)
9. Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A., Beckett, L.: Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's and Dementia 1(1), 55–66 (2005)
10. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. Int. J. of Med. Inform. 73(1), 1–23 (2004)
11. Napel, S.A., Beaulieu, C.F., Rodriguez, C., Cui, J., Xu, J., Gupta, A., Korenblum, D., Greenspan, H., Ma, Y., Rubin, D.L.: Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results. Radiology 256(1), 243–252 (2010)
12. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imag. 18(8), 712–721 (1999)
13. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. on Patt. Anal. and Mach. Intell. 22(8), 888–905 (2000)
14. Simonyan, K., Zisserman, A., Criminisi, A.: Immediate Structured Visual Search for Medical Images. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 288–296. Springer, Heidelberg (2011)
15. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: IEEE Int. Conf. on Comp. Vis., pp. 1470–1477. IEEE Press, New York (2003)
16. Varma, M., Zisserman, A.: Texture classification: Are filter banks necessary? In: IEEE Int. Conf. on Comp. Vis. and Pat. Rec., vol. 2, pp. 691–698. IEEE Press, New York (2003)
17. Weinberger, K., Saul, L.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. 10, 207–244 (2009)

The Synergy of 3D SIFT and Sparse Codes for Classification of Viewpoints from Echocardiogram Videos

Yu Qian[1], Lianyi Wang[2], Chunyan Wang[2], and Xiaohong Gao[1]

[1] School of Engineering and Information Sciences,
Middlesex University, NW4 4BT, U.K.
{y.qian,x.gao}@mdx.ac.uk
[2] Heart Center, First Hospital of Tsinghua University, China
lywang@mail.tsinghua.edu.cn

Abstract. Echocardiography plays an important part in diagnostic aid in cardiology. During an echocardiogram exam images or image sequences are usually taken from different locations with various directions in order to comprehend a comprehensive view of the anatomical structure of the 3D moving heart. The automatic classification of echocardiograms based on the viewpoint constitutes an essential step in a computer-aided diagnosis. The challenge remains the high noise to signal ratio of an echocardiography, leading to low resolution of echocardiograms. In this paper, a new synergy is proposed based on well-established algorithms to classify view positions of echocardiograms. Bags of Words (BoW) are coupled with linear SVMs. Sparse coding is employed to train an echocardiogram video dictionary based on a set of 3D SIFT descriptors of space-time interest points detected by a Cuboid detector. Multiple scales of max pooling features are applied to representat the echocardiogram video. The linear multiclass SVM is employed to classify echocardiogram videos into eight views. Based on the collection of 219 echocardiogram videos, the evaluation is carried out. The preliminary results exhibit 72% Average Accuracy Rate (AAR) for the classification with eight view angles and 90% with three primary view locations.

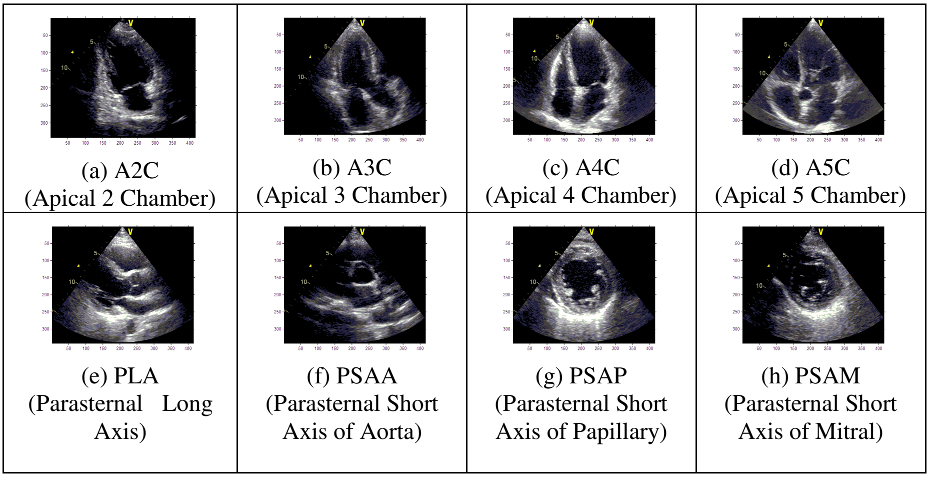
Keywords: Classification of Echocardiogram Video, Cuboid Detector, 3D SIFT, Sparse Coding, SVM.

1 Introduction

Echocardiography remains an important diagnostic aid in cardiology and relies ultrasonic techniques to generate both single image and image sequences of the heart, providing cardiac structures and their movements as well as detailed anatomical and functional information of the heart. In order to capture different anatomical sections of a 3D heart, eight standard views are usually taken from an ultrasound transducer at the three primary positions, which are Apical Angles (AA) (location 1 with 4 view

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 68–79, 2013.

angles), Parasternal Long Axis(PLA) (location 2 with 1 view angle) and Parasternal Short Axis (PSA) (location 3 with 3 view angles) respectively. Example images of these eight views of the 3 primary locations can be seen in Figure 1. The major anatomical structures such as left ventricle are then manually delineated and measured from different view images to further analyze the function of the heart. Hence, the echocardiogram view recognition is the first step for echocardiogram diagnosis.


(a) A2C (Apical 2 Chamber)

(b) A3C (Apical 3 Chamber)

(c) A4C (Apical 4 Chamber)

(d) A5C (Apical 5 Chamber)

(e) PLA (Parasternal Long Axis)

(f) PSAA (Parasternal Short Axis of Aorta)

(g) PSAP (Parasternal Short Axis of Papillary)

(h) PSAM (Parasternal Short Axis of Mitral)

Fig. 1. Eight views of echocaridogram videos

With the advances of the techniques in computer vision, computer-aided echocardiogram diagnosis is becoming increasingly beneficial in recent years, the view also shared in [1,2,3,4]. Their work mainly focuses on spatial and motion representations for the major anatomical structures that can then in turn be used to conduct higher level disease discrimination and similarity search. On the other hand, due to the image variations in the same anatomical structure under different views, prior knowledge of the viewpoint is needed before the treatment on both model selection (i.e. Active Shape Models (ASMs) [2,3]) and filtering (i.e. Edge filter [4]) process. As a result, similar to a clinical workflow, the automatic echocardiogram view classification is the first and essential step in a computer-aided echocardiogram diagnosis system. A number of progresses have been made so far. For example, the work started in [5] indexes echocardiogram videos according to their viewpoint, the work has been subsequently followed by [6,7,8,9]. In [5,8], image-based methods are employed with the focus on the detection of multiple objects and their spatial relationships in an image/frame (e.g. 4 chambers of the heart in A4C). [6,7,9] add motion information in their research. In [9], the features are extracted by calculating magnitude of the gradients in space-time domain of videos whereby a hierarchical classification scheme is performed to reduce the number of misclassifications among the super-classes. In [6], the extraction of motions is conducted by tracking Active Shape Models (ASMs) through a heart cycle that is then projected into an eigen-motion feature space of the viewpoint class for matching. In [6.9], the evaluation are performed only on four views, including Apical 2 Chamber (A2C), Parasternal Long Axis (PLA), Parasternal Short Axis of Papillary (PSAP) and Parasternal Short Axis of Aorta (PSAA) as described in [6], whereas in [9], another four views, which are Apical 4 Chamber (A4C), Apical 2 Chamber (A2C), Parasternal Long Axis (PLA) and Parasternal Short Axis (PSA), are looked at. Additionally, the work specified in [7] utilizes the technique of scale invariant features extracted from the magnitude image that has undergone edge filtered motion as well as Pyramid Matching Kernel (PMK) based on the Support Vector Machine (SVM) for view classification, which has resulted in 81% Average Accuracy Rate (AAR) over a collection of 113 videos with eight views.

In this study, according to the datasets of video clips we collected which consisted of eight viewpoints, we adopt a slightly different approach by utilizing the Bag of Word (BoW) paradigm that is integrated with linear SVMs. Unlike the traditional BoW paradigm [10], sparse coding [11] is employed in this paper instead of Vector Quantization (VQ) to train a video dictionary based on a set of 3D SIFT (Scale Invariant Feature Transform) descriptors of space-time interest points detected by Cuboid detector. Furthermore, instead of using histograms, multiple scales of max pooling features are applied as the representations of echocardiogram videos. Subsequently, the linear multiclass SVMs is employed to classify these echocardiogram videos into eight view groups.

The remaining of this paper is structured as follows. Section 2 explains the methods employed in the study, whist Section 3 shows the experimental results. Conclusion and discussion are drawn in Section 4, which is followed by the sections of Acknowledgment and References.

2 Methodology

Figure 2 schematically illustrates a framework of Bag of visual Word of SVM for the classification of echocardiogram video views, which constitutes visual dictionary generation via sparse coding (left rectangular, coloured in green), video representations based on space-time max pooling of 3D SIFT sparse codes (middle, in red) and echocardiogram video view classification based on multiclass SVM (right, in blue). A codebook of videos is firstly constructed by following the BoW paradigm using 3D SIFT for the feature description of space-time interest points that have been detected using Cuboid detector in advance. Then sparse coding for visual dictionary (a codebook) training starts. Based on a trained codebook, the 3D SIFT of those space-time interest points detected in each video clip are then coded using these codes. The adoption of space-time max pooling of 3D SIFT sparse codes then takes place as echocardiogram video representations. As a result, the classification of video clips is performed using multiclass linear SVMs. The detailed methodology is further accounted for in the next section.

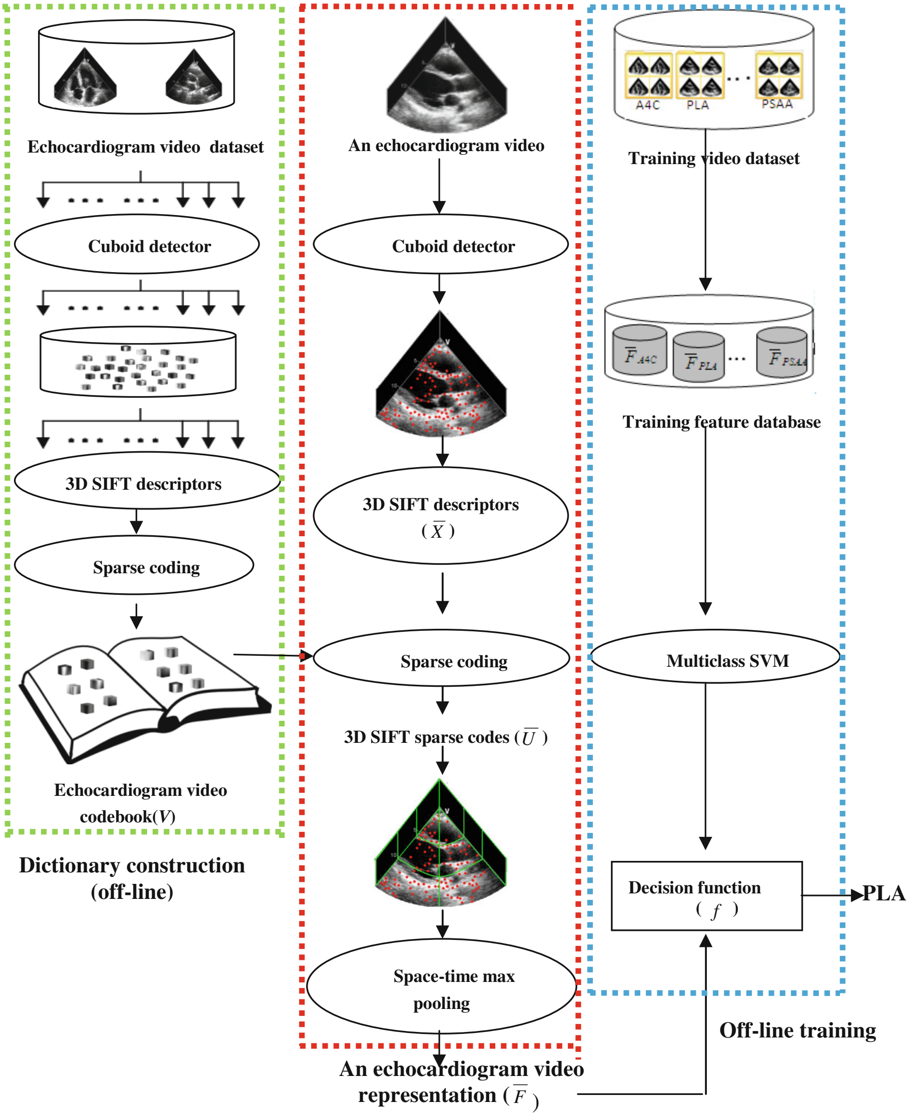


Fig. 2. A framework of bag of word (BoW) SVM recognition

2.1 The Creation of an Echocardiogram Video Codebook

1) Space-Time Interest Point Detector --- Cuboid Detector

A variety of methods exist to detect Space-Time Interest Point (STIP) in image sequences. Typically, STIPs are figured out via firstly calculating a response function over the spatiotemporal locations and scales of image sequences and then by selecting the local maxima of the response function. The evaluation of STIP methods overall the standard video datasets (i.e. KTH actions[1], UCF sports[2], Hollywood2 movies[3] and FeEval[4]) [12,13] have demonstrated our choice of Cuboid detector + 3D Histogram of Oriented Gradients (HOG3D) descriptor that gives better performance in action recognition. In comparison with Harris3D [14] and Hessian3D [15], Cuboid detector [16] overcomes the lacks of temporal response by dealing with temporal data separately with Gabor filters, which not only measures local changes in the temporal domain, but prioritizes the repeated events of a fixed frequency such as heartbeat in echocardiogram video.

The Cuboid detector is a set of separable linear filters with 2D spatial Gaussian smooth kernel and 1D temporal Gabor filters, as such a response function is formulated as

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $I(x, y, t)$ refers to an image sequence; $g(x, y; \sigma)$ the 2D spatial Gaussian smoothing kernel with spatial scale σ , whereas $h_{ev}(t; \tau, \omega)$ and $h_{od}(t; \tau, \omega)$ defined as Eq. (2) are a quadrature (cosine and sine) pair of 1D temporal Gabor filters with temporal scale τ with $\omega = 4/\tau$. Like [12], the scale parameter $\sigma = 2$ and $\tau = 4$ are selected in this study.

$$\begin{aligned} h_{ev}(t; \tau, \omega) &= -\cos(2\pi\omega t) e^{-\frac{t^2}{\tau^3}} \\ h_{od}(t; \tau, \omega) &= -\sin(2\pi\omega t) e^{-\frac{t^2}{\tau^3}} \end{aligned} \tag{2}$$

As a result, space-time interest points are extracted by calculating the local maxima of the response function *R*.

2) Space-Time Interest Point Descriptor --- 3D SIFT Descriptor

After the affirmation of space-time interest points, the representation of these points follows for the further processing. These descriptors should capture space-time neighborhoods of the detected interest points and are usually formulated by using

[1] http://www.nada.kth.se/cvap/actions/

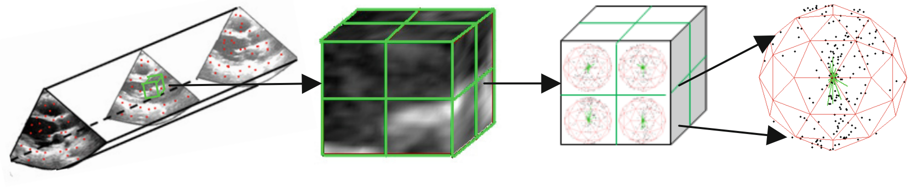
[2] http://vision.eecs.ucf.edu/

[3] http://www.di.ens.fr/~laptev/download.html

[4] http://www.feeval.org/Data-sets/FeEval.html

image measurements such as Histogram of space-time Oriented Gradients (HOG3D) [17], concatenation of Histogram of spatial Oriented Gradients and motion Optical Flow (HOG/HOF) [18], and 3D Speeded Up Robust Feature (SURF3D) [15]. According to our previous study, 3D SIFT, also known as HOG3D gives robust feature description and is therefore employed in this study to describe visual feature of space-time interest points detected by Cuboid detector.

As shown in Figure 3 (a and b), the 12 x 12 x 12 neighbourhood volume around an interest point is selected and then divided into 2x 2 x2 = 8 sub-volumes. For each sub-volume, the gradient magnitude and orientation of each voxel in the sub-volume are calculated by using Haar wavelet transform along x, y and z direction respectively, and then the magnitude of the gradient is accumulated to the corresponding bin of the gradient orientation. The tessellation based orientation histogram [19] is then implemented in this study. By using the tessellation technique, each bin of 3D gradient orientation is approximated with a mesh of small piece of 3D volume seen as a triangle in Figure 3(d). The gradient orientations pointing to the same triangle then belong to the same bin, as marked by the black points in Figure 3(d). The total number of the bins is calculated as 20 x (4 ^ Tessellation level). The Tessellation level decides the number of constituting triangle surfaces, i.e., the number of bins of gradient orientation in 3D space. In this study, the Tessellation level is set to 1, thus resulting in 80 bins. Each sub-volume is accumulated into its own sub-histogram. Subsequently, the 3D SIFT descriptor *X* of each interest point is of 2 x 2 x 2 x 80 (= 640) dimensions.



(a) An echocardiogram video sequence

(b) Neighborhoods of a space-time interest points

(c) 3D SIFT descriptors(*X*) of a space-time interest points

(d) Histogram of gradient orientation

Fig. 3. 3D SIFT descriptors

3) Echocardiogram Video Vocabulary Construction---Sparse Coding

Once the 3D SIFT features are extracted from each space-time interest point, which are considered as candidates for unit elements, or the “words” in the visual dictionary, sparse coding is employed.

Sparse coding [11] that models data vector as a sparse linear combination of a set of basic elements called dictionary is applied to construct visual dictionary and encodes each descriptor of an image by solving an optimization problem as formulated in Eq.(3).

$$\min_{U,V} \sum_{m=1}^{M} \left\| x_m - Vu_m \right\|_2^2 + \lambda \left\| u_m \right\|_1 \tag{3}$$

$$\textit{Subject to:} \quad \left\| v_i \right\| \leq 1, \qquad \forall\, i = 1,\ldots, K$$

Where $X = \left[x_1, x_2, \ldots x_M \right] \left(x_m \in R^{dx1} \right)$ represents a set of 3D SIFT descriptors from echocardiogram video dataset; $V = \left[v_1, v_2, \ldots v_K \right] \left(v_i \in R^{dx1} \right)$ refers to the K bases, called the dictionary or codebook; $U = \left[u_1, u_2, \ldots u_M \right] \left(u_m \in R^{Kx1} \right)$ remains the sparse codes for video based on codebook V, and λ is the coefficient to control the amount of L_1 norm ($\left\| \cdot \right\|_1$) regularization.

In the training stage, 80000 interest points as the training data set are randomly selected from all interest points in our video clips, and their 3D SIFT descriptors are applied to off-line training on the codebook V with the size of K = 4000 by solving Eq.(3) using alternating optimizing technique over V or U while fixing the others.

2.2 Echocardiogram Video Representations --- Space-Time Max Pooling of 3D SIFT Sparse Codes

In the coding stage, 3D SIFT descriptors x_i extracted from each interest point can be encoded as u_i by inputting the trained codebook V in Eq. (3). A clip of video is then described as a set of 3D SIFT sparse codes $\overline{U} = \left[u_1, u_2, \ldots u_N \right]$, where N is the total number of the interest points in the video.

In order to describe the local visual features, a video is divided into a number of sub-volumes as illustrated in Figure 4. According to the characteristics of our dataset that lacks heartbeat ECG data, the alignment with time scale is unavailable. As a direct result, although a group of videos belonging to the same view might have been captured from the similar locations and angles, they can be recorded at different starting times of a heartbeat circle, implying two interest points from two different videos being not comparable while in the time domain. Therefore, the grouping of these videos is only fulfilled in the space domain (along horizontal and vertical direction), instead of time domain. In this study, a video clip is divided into 3 sub-volumes in the geometric space of space-time (Up, Middle and Bottom) with equal distance along vertical direction and 2 sub-volumes (Left and right) along a vertical center plane respectively as shown in the middle graph of Figure 4, and then is further divided into 6 sub-volumes as shown in the right of Figure 4. In total, 12 (=1+3+2+6) sub-volumes are created in this way to reflect different scales.

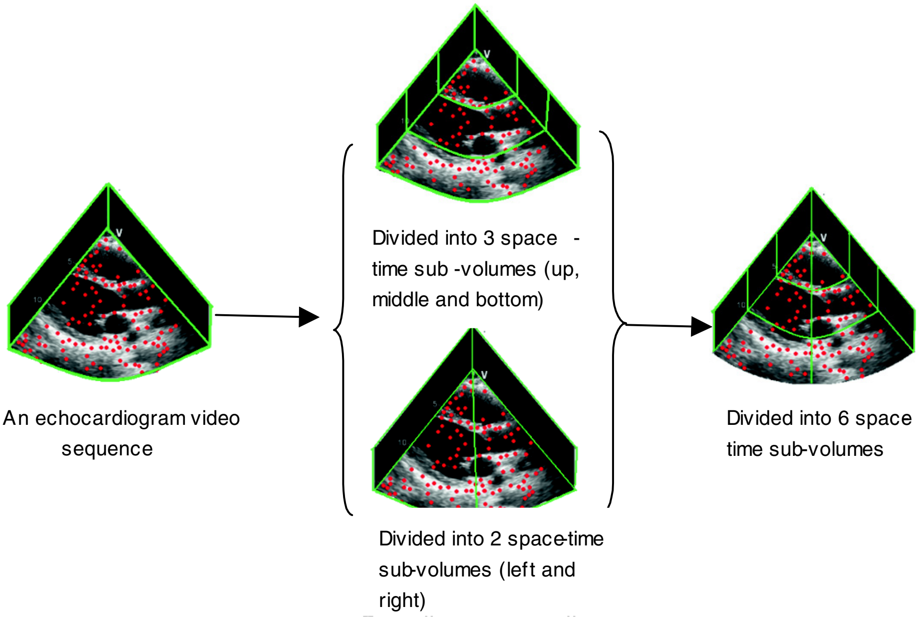

Divided into 3 space - time sub -volumes (up, middle and bottom)
An echocardiogram video sequence
Divided into 6 space time sub-volumes
Divided into 2 space-time sub-volumes (left and right)

Fig. 4. Space-time max pooling

Similar to [11], the representations for each sub-volume noted as $F = \{f_i, i = 1,2,\cdots,K\}$ are computed by a max pooling function as given below in Eq. (4).

$$f_i = \max\{|u_{1i}|, |u_{2i}|, \cdots, |u_{Si}|\} \tag{4}$$

Where K indicates the size of the codebook V . In Eq. (4), S refers to the total number of the interest points in the sub-volume. The pooled features from all the sub-volumes at different spatial levels are then concatenated to form a space-time representation of a video $\overline{F} = \{F_j, j = 1,2,\cdots,P\}$, where P = 12 is the total number of sub-volumes in a video clip.

2.3 Echocardiogram Video Classification --- Linear SVMs

Following the pooling of sub-volume features, the classification of video clips is performed using a multiclass SVM with a linear kernel as formulated in Eq. (5).

$$k(\overline{F_i}, \overline{F_j}) = \overline{F_i}^T \overline{F_j} \tag{5}$$

Where $\overline{F_j}$ is the feature representation of video *j*. With regard to binary classification, an SVM aims to learn a decision function based on the training dataset as defined in Eq. (6).

$$f\left(\overline{F}\right)=\sum_{i=1}^{n} a_i k\left(\overline{F}_i,\overline{F}\right)+b \tag{6}$$

In order to obtain an extension to a multi-class SVM, the trained videos are represented as $\left\{\left(\overline{F}_i,l_i\right)\right\}_{i=1}^{n}$, where $l_i \in \{1,2...L\}$ denotes the class label of trained video *i*. One-against-all strategy is applied to train the total number of L binary classifiers.

3 Experimental Results

3.1 Dataset

In this paper, a total of 219 echocardiogram videos are collected from 72 different patients (containing 14 wall motion abnormalities and 58 normal cases) in the First Hospital of Tsinghua University, China. All videos are captured with duration of 1 second from GE Vivid 7 or E9 and are stored in the DICOM (Digital Imaging and Communications in Medicine) format with the size of 434 pixel x 636 pixel x 26 frame. Each clip belongs to one of the eight different views, as detailed in Table 1. The ground truth data of eight different view videos is created by clinicians in the Heart Center of the First Hospital of Tsinghua University.

Table 1. Classes in the Dataset

View	A2C	A3C	A4C	A5C	PLA	PSAA	PSAP	PSAM	Total
Videos	42	32	34	7	37	39	19	9	219

3.2 Experiment and Results

In order to train an echocardiogram codebook, 80,000 interest points are randomly selected from all interest pointes in 219 video clips, and their 3D SIFT descriptors yield a feature database with the size of 80,000 (number of trained interest points) x 640 (size of 3D SIFT descriptors), which are then subsequently applied to train a codebook with the size of 4000 (size of the codebook) x 640 (size of 3D SIFT descriptors) using the approach of sparse coding with 10 iterations. Based on the trained codebook, all interest points from the 219 videos are represented by the 3D SIFT sparse codes. A space-time max pooling is subsequently applied to obtain video representations with the size of 4000 (size of codebook) x 12 (sub-volumes). Due to the small dataset in this study, we employ the leave-one-out methodology, i.e., when testing a video clip, the entire dataset exclude test video is used for SVM training. The classification results for the eight views are visualized in a confusion matrix as shown in Table 2.

Table 2. Confusion matrix for 8 echocardiogram view classification

		Classification Results								**Accuracy Rate (AR)**
		A2C	A3C	A4C	A5C	PLA	PSAA	PSAP	PSAM	
Ground Truth	A2C	32	2	6	0	0	2	0	0	0.76
	A3C	6	17	6	0	0	3	0	0	0.53
	A4C	5	1	26	0	2	0	0	0	0.76
	A5C	1	0	2	4	0	0	0	0	0.57
	PLA	1	0	0	0	34	2	0	0	0.92
	PSAA	2	0	0	0	4	28	1	4	0.72
	PSAP	0	0	0	0	2	5	12	0	0.63
	PSAM	0	0	0	0	1	3	1	4	0.44
Error Rate (ER)		0.32	0.15	0.35	0	0.21	0.3	0.14	0.5	

The values in the last column of Table 2 are Accuracy Rate (AR) values for each class, whereas the values in the last row refer to Error Rate (ER) for each class. In summary, the average AR (AAR) for all classes is 72% (157/219), and the average ER (AER) is 28% (62/219). According to the data in Table 2, the most erroneous classification takes place within the classes having the similar view points, such as views taken from Apical angles (4 views) and Parasternal Short Axis (3 views). The unique view of PLA gives the best performance (AR=92%).

Our method is also tested on three primary view locations taken from Apical angles (including A2C, A3C, A4C and A5C, with 115 datasets in total), Parasternal Long Axis (PLA, with 37 data) and Parasternal Short Axis (including PSAA, PSAP and PSAM, with 67 data in total). The classification results are shown in Table 3. The AAR for the three classes is 90% (197/219), and the AER is 10% (22/219), suggesting the significant benefit of proposed synergy.

Table 3. Confusion matrix for 3 primary view locations

		AA (Apical Angle)	PLA (Parasternal Long Axis)	PSA (Parasternal Short Axis)	**Accuracy Rate (AR)**
Ground Truth	AA	112	0	3	0.97
	PLA	3	31	3	0.84
	PSA	9	4	54	0.81
Error Rate (ER)		0.1	0.11	0.1	

4 Conclusion and Discussion

Due to the lack of ECG data in our datasets, comparison with the similar work as addressed at [7] might not be straitforward if not possible. In their study, data alignment is performed first to ensure all the video data starting from the same heartbeat cycle, whereas in our case, this alignment in the time domain is omitted via using space-time max pooling for feature representations (detailed in Section 2.2), making our appraoch more challenge. In addition, their AAR value of impressive 81% is based on 113 videos, whareas ours of 72% of AAR arises from 219 clips. All in all, each approach has both pros and cons and is usually talored based on the characteristics of each data collection. Therefore the future work includes cross evaludation given the availability of different datasets.

This papre presents that the synergy of the well-known alrgorithms obtained in each individual computer vision field can be possible to produce an improved results in a clinical sector. In dealing with echocardiographies, challenges remain on not only the low resolution that an ultrasonic image endures but also the computational complexity and time cost while processing video images. With the availabilty of ECG data in the future, the calibration of time scale can be achieved, which however might introduce extra porcessing cost. The future work also include the inclusion of larger datasets to further varify the proposed synergy.

Acknowledgments. This research forms part of WIDTH project that is financially funded by EC under FP7 programme with Grant number of PIRSES-GA-2010-269124. Their support is gratefully acknowledged. The authors would also like to express their gratitude to one of our PhD students, Dr. Rui Hui, who has spent considerable time in helping the classification of training datasets based on his in-depth clinical knowledge.

References

1. Syeda-Mahmood, T., Wang, F.: Characterizing Normal and Abnormal Cardiac Echo Motion Patterns. In: Computers in Cardiology, pp. 725–728 (2006)
2. Syeda-Mahmood, T., Wang, F., Beymer, D., London, M., Reddy, R.: Characterizing Spatio-temporal Patterns for Disease Discrimination in Cardiac Echo Videos. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 261–269. Springer, Heidelberg (2007)
3. Beymer, D., Syeda-mahmood, T.: Cardiac Disease Detection in Echocardiograms Using Spatio-temporal Statistical Models. In: Annual Conference of IEEE Engineering in Medicine and Biology Society, EMBS (2008)
4. Kumar, R., Wang, F., Beymer, D., Syeda-mahmood, T.: Cardiac Disease Detection from Echocardiogram using Edge Filtered Scale-Invariant Motion Features. In: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA (2010)

5. Ebadollahi, S., Chang, S.F., Wu, H.: Automatic View Recognition in Echocardiogram Videos Using Parts-based Representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2–9 (2004)
6. Beymer, D., Syeda-Mahmood, T., Wang, F.: Exploiting Spatio-temporal Information for View Recognition in Cardiac Echo Videos. In: IEEE Workshop on Mathematical Methods in Biomedical Imaging Analysis (MMBIA), pp. 1–8 (2008)
7. Kumar, R., Wang, F., Beymer, D., Syeda-mahmood, T.: Echocardiogram View Classification Using Edge Filtered Scale-invariant Motion Features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 723–730 (2009)
8. Zhou, S.K., Park, J.H., Georgescu, B., Simopoulos, C., Otsuki, J., Comaniciu, D.: Image-based Multiclass Boosting and Echocardiographic View Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1559–1565 (2006)
9. Otey, M.E., Bi, J., Krishnan, S., Rao, B., Stoeckel, J.: Automatic View Recognition for Cardiac Ultrasound Images. In: Workshop on Computer Vision for Intravascular and Intracardiac Imaging, pp. 187–194 (2006)
10. Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: IEEE Conference on Computer Vision (ICCV), pp. 1470–1477 (2003)
11. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1794–1801 (2009)
12. Wang, H., Ullah, M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of Local Spatio-temporal Features for Action Recognition. In: British Machine Vision Conference (BMVC), pp. 127–137 (2009)
13. Stöttinge, J., Goras, B., Sebe, N., Hanbury, A.: Behavior and Properties of Spatio-temporal Local Features under Visual Transformations. In: ACM International Conference on Multimedia (ACMMM), pp. 1155–1158 (2010)
14. Laptev, I.: On Space-time Interest Points. IEEE International Journal on Computer Vision (IJCV), 107–123 (2005)
15. Willems, G., Tuytelaars, T., Van Gool, L.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
16. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-temporal Features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), pp. 65–72 (2005)
17. Kläser, A., Marszałek, M., Schmid, C.: A Spatio-Temporal Descriptor Based on 3D Gradients. In: British Machine Vision Conference (BMVC), pp. 995–1004 (2008)
18. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning Realistic Human Actions from Movies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8 (2008)
19. Scovanner, P., Ali, S., Shah, M.: A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition. In: ACM Conference on Multimedia, pp. 357–360 (2007)

Assessing the Classification of Liver Focal Lesions by Using Multi-phase Computer Tomography Scans

Auréline Quatrehomme[1,2], Ingrid Millet[3], Denis Hoa[1], Gérard Subsol[2], and William Puech[2]

[1] IMAIOS, Montpellier, France
aureline.quatrehomme@imaios.com
[2] LIRMM, Université Montpellier 2 / CNRS, Montpellier, France
[3] Department of Medical Imaging, CHU Lapeyronie, Montpellier, France

Abstract. In this paper, we propose a system for the automated classification of liver focal lesions of Computer Tomography (CT) images based on a multi-phase examination protocol. Many visual features are first extracted from the CT-scans and then labelled by a Support Vector Machine classifier. Our dataset contains 95 lesions from 5 types: cysts, adenomas, haemangiomas, hepatocellular carcinomas and metastasis. A Leave-One-Out cross-validation technique allows for classification evaluation. The multi-phase results are compared to the single-phase ones and show a significant improvement, in particular on hypervascular lesions.

Keywords: Medical Imaging, Computer Aided Diagnosis, Liver focal lesions, Multi-Phase Computer Tomography, Classification.

1 Introduction

Computer Aided Diagnosis (CAD) is a current dynamic field of research, with the help of recent imaging device improvements. For example, by integrating computer assistance in the diagnosis process of liver lesions, we can improve the efficiency of medical expertise and accuracy in classifying, detecting or segmenting the liver lesions. In this paper, we describe a preliminary study of a new method to classify hepatic lesions, without any detection or segmentation (as described in [1]), which is based on 4-phase CT imaging.

Section 2 introduces research on liver CT Computer Aided Diagnosis and some references dealing with multi-phase scans. Section 3 describes precisely how our dataset was built. Section 4 presents the method used and the results are analyzed in Section 5. In Section 6, we present some perspectives to improve these first results.

1.1 Multi-phase CT Acquisition

X-ray CT captures a large series of two-dimensional x-ray images, taken around one single rotation axis. Its usage has dramatically increased over the last two

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 80–91, 2013.

decades, in particular for abdominal exploration. In order to improve the contrast of the captured images, and therefore the accuracy of the diagnosis, contrast media injection is widely used. One series is first captured on the patient (pre-injection phase). The patient then receives the injection, and 3 series are taken at three different times: the first one, just after the injection, is called the arterial phase. The second, a few tens of seconds later, the portal phase. The last one, a few minutes after the injection: the late phase.

The diffusion of the media over the different phases captured will enhance the vessels and lesions. Radiologists would not imagine making a diagnosis without the essential temporal information provided from these multiphase scans. Indeed, the contrast enhancement varies from one phase to another: a lesion indistinguishable from the healthy liver in one phase will be revealed in another phase. This property is illustrated in Table 1, which visually shows these variations. Moreover, different types of lesions have different enhancement patterns and timelines. We have summarized information from a paper on strategies for hepatic CT and MRI imaging [1] in Table 2.

1.2 CT Liver Lesion Classification

Various papers have been published on Computer Aided Diagnosis (CAD) using liver CT scans. A team from Stanford focused on the shape of 8 types of liver nodules in [2], while they added in [3] semantic features to texture and boundary features in order to distinguish cysts, hemangiomas and metastases. These two papers apply their methods to Content-Based Image Retrieval (CBIR), which returns the images of the database which are the most similar to the query image. Mougiakakou *et al.* [4] applied multilayer perceptron neural networks, as

Table 1. Visual appearance of lesions by type and phase illustrating the importance of multi-phase CT scans

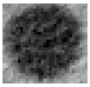
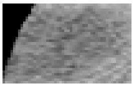
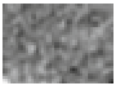
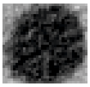
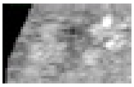
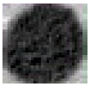
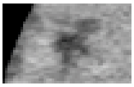
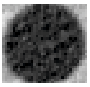
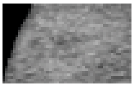
Phase \Lesion	**C**yst	**A**denoma	**H**aemangioma	**H**CC	**M**etastasis
1 pre-injection					
2 arterial phase					
3 portal phase					
4 late phase					

Table 2. CT scan scenario and their clinical context: captured phases and context

CT scan scenario	Clinical context
Single-phase (portal)	No suspicion of a specific hepatic pathological condition
Dual-phase (arterial, portal)	Disease scenario with the primary cause outside of the liver, hypervascular hepatic metastases suspected
Triple-phase(before injection, arterial, portal)	Known or suspected cirrhosis, HCC, FNH or adenoma

well as a combination of primary classifiers, to the classification of liver tissue into healthy liver, cyst, hemangioma and hepatocellular carcinoma.

Surprisingly though, most image databases found in the literature contain images from one single CT phase despite its importance in the diagnosis process. We found two attempts for the study of the multi-phase CT classification of liver lesions, which will be presented below.

Duda *et al.* [5] focus on texture characteristics. Their database contains 165 lesions from 3-phase CT acquisition (no contrast, arterial and portal phase). They tested 4 sets of features (First-Order statistics, Law entropy, Run-Length matrix features and Co-occurrence matrix measures) independently at each phase, before all sets of features at each phase, then each feature set at all phases, finally all features at all phases altogether. SVM and the Dipolar Decision Tree were both used as classifiers to distinguish between healthy liver, HCC and cholangiocarcinoma.

Ye *et al.* [6] compared the results obtained from Support Vector Machines (SVM) classification on each phase with textural features: first order statistics as well as statistics computed over the image co-occurrence matrix. Furthermore, they introduced temporal tendency features over the phases. Their database consists of 131 four-phase examinations. The study is carried out on 4 classes: healthy liver, cyst, HCC and haemangioma, and the classification is always binary: normal vs. abnormal, cyst vs. other diseases, haemangioma vs. HCC. The temporal features idea seems interesting, although its application here is quite limited as the different features are computed over the mean value of the pixels (heterogeneous lesions might be hard to distinguish in this case). We also regret the lack of classification of the values obtained on the four phases, and the limitations resulting from by the binary classification scheme.

2 Data

2.1 Database Construction

With the help of 2 radiologists, we opted for five lesion diagnosis classes: cysts, metastasis and hypervascular lesions: adenomas, haemangiomas and hepatocellular carcinoma (HCC) which are presented in Table 1 and Table 3. This set of diagnosis types cover the majority of focal hepatic lesions. Cysts are both benign and very commonly observed, but as their texture is homogeneous and their contours well defined, they have been under-represented in our database. On the other hand, adenomas, which are very rare but heterogeneous, are more present than in clinical reality. The repartition of the lesion types in our database is presented in Table 3. Our objective is not to determine whether the liver is in good condition or unhealthy, but to distinguish between nodular hepatic lesions, so no healthy tissue is present in the database.

Table 3. Lesion class repartition in our database

CLASS	Cysts	Adenomas	Haemangiomas	HCC	Metastasis	TOTAL
NUMBER	25	10	9	13	38	95

This is a retrospective analysis of daily CT scans conducted on two different scanners at the University Hospital of Montpellier between 2008 and 2011, so no patients were irradiated for our research, and no particular procedure other than the routine protocol was followed for the capture. An experimented radiologist looked for particular diagnosis clinical cases, and analyzed the CT images as well as the reports and complementary histological results which confirmed the diagnosis.

95 lesions of 40 different patients were selected to constitute our database. Its size is comparable to those of similar studies [5,6]. The slice thickness and the number of phases vary, depending on what the radiologist was interested to see in the examination, which therefore determined the protocol. Slice thickness goes from 1.25 to 3 millimeters. 16 cases contain two phases images, 7 cases three phases, and 78 the four phases.

2.2 Data Pre-processing

We work directly with the DICOM images. As the pixel values of this format represent tissue densities, the entire range of the scale is kept and the grey levels are not normalized. The lesions are present on several CT slices, therefore a 2D rectangular bounding box was drawn around the lesions by an experimented radiologist in the middle single slice. No precise segmentation was done, in order to avoid certain problems, in particular due to the irregularity of the contours. In order to refine this rectangular box, and because we are working on *focal* lesions, the bounding ellipse in the rectangular zone defined by the radiologist will be used as region of interest (ROI), as presented in Figure 1, in the ”Data acquisition and pre-processing” section. Therefore, lesion tissue will be studied instead of healthy liver. The ROI size ranges from 9*12 to 165*189 pixels, which is representative of the variety of hepatic lesion sizes.

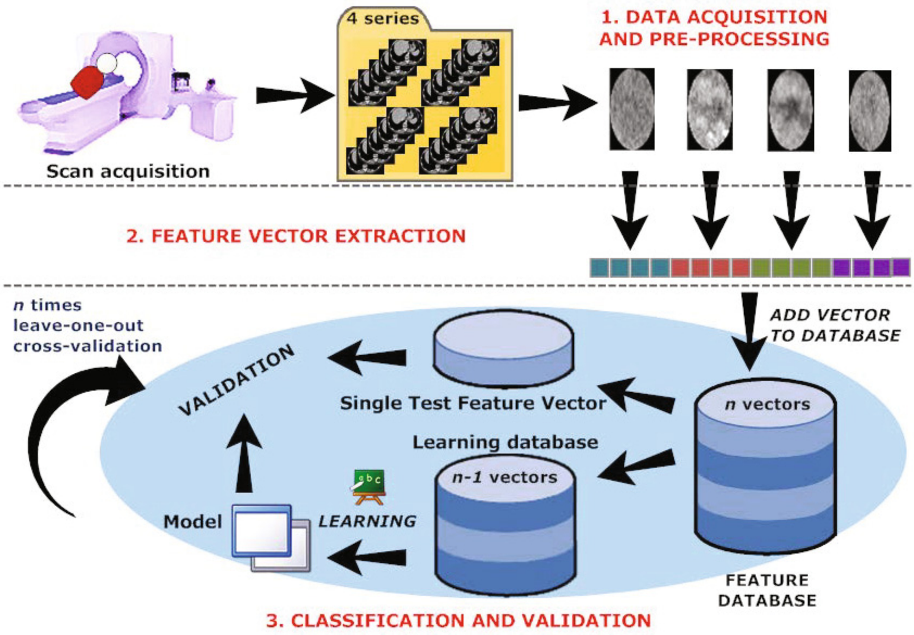


Fig. 1. 3-step framework of proposed system: ROI then visual features extraction, before classification and evaluation

3 Method

3.1 System General Framework

Figure 1 presents an overview of the proposed system. Each lesion is a set of one to four 2D DICOM images, depending on the number of phases captured from the patient, on which a Region Of Interest (ROI) is extracted. Visual features are computed over these images and form multi-phase vectors, which are entered into a Support Vector Machine (SVM) classifier. A Leave-One-Out (LOO) cross-validation technique is finally conducted for classification evaluation.

First, the feature extraction step will be described in section 4.2 , then the classification scheme in section 4.3. As in the papers by Duda *et al.* [5] or Ye *et al.* [6], our framework is broken down into 3 steps: feature extraction, training a classifier and classification (see Table 4 for comparison).

3.2 Feature Extraction

For segmentation, detection, retrieval or classification, the basic principle is to extract some visual features, or descriptors, from images. They describe the characteristics of the image, express its content (grey levels/colours, texture or shape). They are computed on the whole image, on each block obtained by

Table 4. 3 multi-phase system comparison based on multiple data, features and classification criteria

Characteristic	**Ye *et al.* [6]**	**Duda *et al.* [5]**	**Our work**
Lesion number	131	165	95
Lesion size	unknown	unknown	from 9*12 to 165*189 pixels
Phases	4	3 (late phase absent)	– 4-phase: 78 – 3-phase: 7 – 2-phase:16
Diagnosis classes	– HCC – cyst – haemangioma – healthy	– HCC – healthy – cholangiocarcinoma	– HCC – cyst – haemangioma – adenoma – metastasis
Region Of Interest	16x16 pixels square in the lesion manually delineated	manual circle of 30 to 70 pixels radii	manual rectangular bounding box around the lesion then automatically extracted inscribed ellipse
Features	– First Order Statistics, – Co-occurrence matrix statistics, – Temporal features	– First Order Statistics, – Co-occurrence matrix statistics, – Law measures, – Run-Length matrix features	– First Order Statistics, – Gaussian Markov Random Fields, – Law measures, – Unser histograms statistics
Classifier	SVM	SVM	– SVM – Dipolar Decision Tree
Classification	3 binomial sequential classifications: – healthy vs. pathology – if pathological: cyst vs. non cyst – if non-cyst: HCC vs haem.	Distinguish the 3 classes	Distinguish the 4 classes

dividing the image in small equally sized patches, or on Regions of Interest (ROIs), which have been delineated by a manual or automatic segmentation process. A review of the features can be found in [7] for recent CBIR systems, and in [8] for medical image classification.

We decided to begin our study with a few common features computed over the 4 phases, described below. All of them are extracted over the ellipsoid 2D ROI defined in Section 2. The first one, Unser histograms statistics, is an exception as it has never been tested to our knowledge.

Unser Histograms: Unser proposed in 1986 [9] an alternative method to the Grey-Level Co-occurrence Matrix (GLCM) computation, which reduces the memory requirement as well as the calculation time. GLCM, over which Haralick's well-known texture descriptors are computed, is replaced by estimates of the first order probability functions along its principal axes, which correspond to the second order probability functions over the image. These are called sum and difference histograms and they are extracted over four different directions. 9 statistical descriptors are then calculated over these two histograms in each direction, ending up with 36 attributes. Unser claims they are as accurate for classification as the GLCM statistics. We tested both Haralick and Unser measures and ended with similar and even better results with Unser, with the computation advantage already cited.

Law Measures: Kenneth I. Law proposed in 1980 [10] texture energy measures, which have been used for various applications. Its method to extract texture features is carried out in 3 steps. First, 25 convolution kernels are applied to the image. Secondly, a texture energy measure is computed on each convolved pixel by a windowing operation, and a new image is formed. Finally, these energy images are normalized then combined in order to obtain 14 rotation invariant final images. Mean and standard deviation are finally computed over them, ending with 28 attributes.

Gaussian Markov Random Fields Measures: Markov Random Fields systems model the dependency phenomena amongst image pixels using a statistical approach. The main idea is that, while neighboring pixels usually have the same intensity in an image, pixel values are independent of the pixels beyond that area. The image is therefore seen as a sample of a random process, where correlation between pixels is proportional to their geometric separation. Instead of being the real probability function computed over the image pixels, the field is a Gaussian in order to avoid high computational problems. The GMRF measures are its average, its standard deviation and 4 parameters named thetas. We keep standard deviation and thetas, while rejecting its average, which approximates very closely the image grey level average.

Histogram Statistics: mean, standard deviation, skewness and kurtosis computed over the grey-level histogram.

Our final set contains 303 attributes over grey levels and texture, on each phase. The feature vector for each lesion contains all the measures side to side, one phase before another. All feature vectors are pre-computed in order to speed up the system.

3.3 Classification

Weka is a collection of machine learning algorithms, written in Java and developed at the University of Waikato, New Zealand (see [11] for an introduction). It can deal with missing values, which is helpful in our case where each CT scan consist of two to four series.

We tried several implemented processes before setting our choice on a classical method: Support Vector Machine (SVM). The algorithm implementation is called Sequential Minimal Optimization (SMO) and was proposed by John Platt [12]. The Support Vector Machines principle is to separate the data by a hyperplane (or a set of hyperplanes) in a high or infinite-dimensional space. In this new space, separations in the data that could not be seen in the initial one may be revealed.

Before the classification, three pre-processing actions are conducted. First, missing values of each attribute are replaced by its mean. Our feature vectors do not all have the same length, depending on the number of phases of the CT acquisition. Then, nominal attributes are transformed into binary ones. Indeed, the SVM algorithm builds several binary models, one for each pair of classes. Finally, feature measures are normalized. The SVM kernel here is polynomial, with a 1.0 exponent.

3.4 Classification Validation

A Leave One Out (LOO) cross-validation technique is conducted.

Cross-validation is used to estimate how accurately our predictive model will perform in practice. One round of cross-validation consists of partitioning a sample of data into 2 complementary subsets. The analysis is performed on the first one (the training set), while the second one (testing set) is for validation. In order to reduce the effects of variability, multiple rounds as described are performed, using different partitions. The validation results are finally averaged over the rounds. Cross-validation gives more realistic results than classification and validation on the same complete database.

As its name suggests, in LOO cross-validation, a single observation of the set is designated as the validation data, and the remaining observations as the training data. The classification is conducted exhaustively n times, with n the number of observations, such that each one is used once for testing.

This classification with cross-validation is conducted in 0.19 seconds in the case of multi-phase, and 0.06 seconds in the case of mono-phase (for the complete lesion database). We are able to classify new lesions in real-time.

4 Results and Analysis

4.1 Analysis Scheme

The confusion matrix from multi-phase classification results was obtained and compared to the one from the portal phase. We extracted precision (also called true predictive value) and recall (also known as sensitivity) measures, as well as the F-measure of the test. Precision is a measure of the accuracy provided that a specific class has been predicted, whereas recall represents the ability to select instances of a certain class from a dataset. F-measure is an indicator of the global classification accuracy and it is defined by the weighted harmonic mean of precision and recall.

4.2 Precision, Recall and F-measure

The three measures chosen to evaluate our classification can be visualized in Figure 2. The same tendency can be observed over the three bar charts. The weighted average values show a global improvement of the three statistics by the introduction of multi-phase (+12% for precision and recall, +13% for F-score). If we have a closer look at the results obtained for each lesion type, the major phenomenon observed here is the spectacular improvement due to the multi-phase CT acquisition of the three measures for haemangioma and HCC (respectively from 56 to 63% and from 31 to 50%). Adenoma also benefits from multi-phase images, but to a lesser extent (from 5 to 8%). Regarding cysts and metastasis, portal phase evaluation seems sufficient: results are stable on cysts (8% maximum variation), and multi-phase has little positive influence on precision and F-measure (from 7 to 10%), whereas recall values goes down from 19%.

4.3 Confusion Matrices

Regarding the confusion matrix obtained with portal phase feature classification, cysts, adenomas and metastasis are quite well recognized (respectively 22 out of 25, 8 out of 10 and 35 out of 38), whereas heamangiomas and HCC are never recognized. One-third of the heamangiomas (3) have been labelled as adenomas and the other two (6) as metastasis. All HCC have also been classified as metastasis. This mislabelling on single phase analysis is expected as these lesions are hypervascular lesions and may be indistinguishable from a healthy liver at the portal phase. This confusion observed in portal phase has been pointed out, for example in [13], which studied the enhancement patterns of focal liver lesions during arterial time. At this phase, HCC, haemangiomas and metastasis may alltogether present an homogeneous enhancement pattern, HCC and metastasis may both present abnormal internal vessels or variegated, complete ring or no enhancement pattern at all, while haemangiomas and metastasis may both present peripheral puddles or incomplete ring.

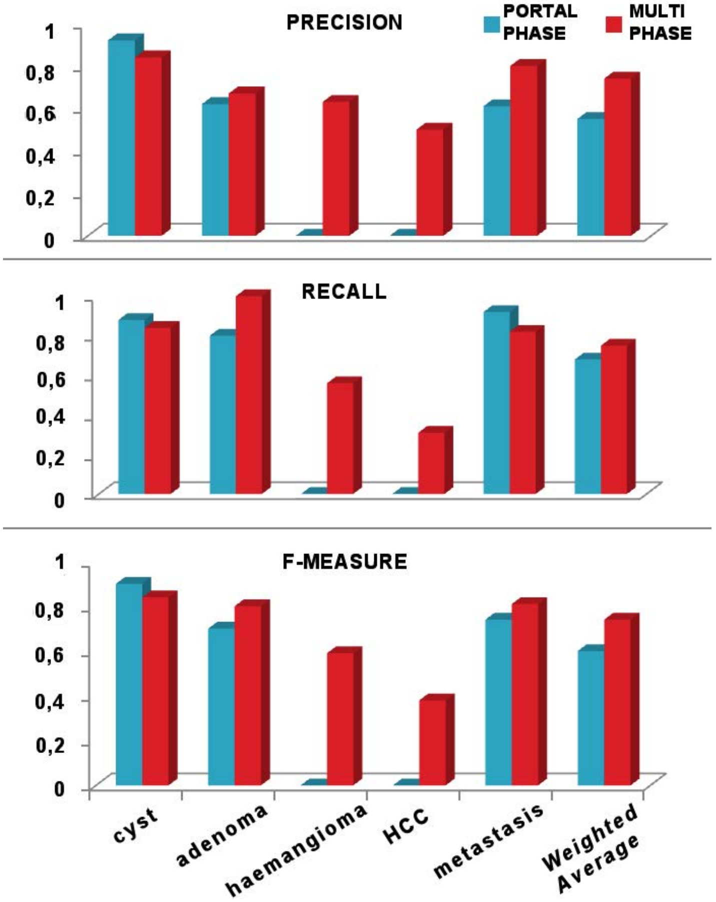


Fig. 2. Precision, Recall and F-measure values obtained on each lesion class as well as on the weighted average of all classes from portal phase and multi-phase classification

	PORTAL PHASE					MULTI-PHASE				
CLASS \FOUND	Cy.	Ad.	He.	HCC	Me.	Cy.	Ad.	He.	HCC	Me.
Cyst	**22**	1	1	0	1	**21**	3	0	0	1
Adenoma	0	**8**	0	0	2	0	**10**	0	0	0
Haemangioma	0	3	**0**	0	6	1	0	**5**	3	0
HCC	0	0	0	**0**	13	0	0	2	**4**	7
Metastasis	2	1	0	0	**35**	3	2	1	1	**31**

Fig. 3. Confusion matrix on classification results: on the left: the real lesion type, on top: the labels determined by the classifier

As regards the confusion matrix obtained with multi-phase feature classification, and compared to portal phase results, HCC and haemangiomas recognition sharply increases, adenomas only slightly, cysts are stable while the metastasis score is falling marginally. What is significant to our earlier remark is that now 3 cysts are seen as adenomas, and the metastasis scheme has spread out over all other diagnosis classes. A sequential two-step classification could be considered: the first one, during the portal phase, to distinguish between cyst, metastasis or other nodule, and the second one, on all phases, if the first classifier labelled the instance as "other", to differentiate between adenomas, haemangiomas and metastasis. This idea coincides with the scheme detailed by Ye *et al.* in their paper [6]. For their part, haemangiomas and HCC are confused with each other in this matrix, and half of the HCC are still confused with metastasis as in the portal phase classification.

5 Conclusion

This paper presents a classical approach for liver lesion classification applied on multi-phase CT scans on the contrary of a majority of other studies which are based on the portal phase only. In this manner, the contrast enhancement patterns of the hepatic lesions can be taken into account.

We applied our system to a database of 95 2D CT images from 40 patients and evaluated its performances and compared them by using the portal phase only. The experimental results show a significant improvement of the classification results by using multi-phase scans, in particular for heamangiomas and HCC lesions. It is important to underline that we work on five diagnosis classes which spans most of the cases of liver lesions.

In the future, we plan to study the influence of each feature on the classification results in order to propose an automated feature selection. Temporal changes among the phases as well as a classification in sequence seem interesting leads to follow.

References

1. Boll, D.T., Merkle, E.M.: Diffuse Liver Disease: Strategies for Hepatic CT and MR Imaging. RadioGraphics 29, 1591–1614 (2009)
2. Xu, J., Faruque, J., Beaulieu, C.F., Rubin, D., Napel, S.: A Comprehensive Descriptor of Shape: Method and Application to Content-Based Retrieval of Similar Appearing Lesions in Medical Images. Journal of Digital Imaging, 1–8 (2011)
3. Napel, S., Beaulieu, C., Rodriguez, C., Cui, J., Xu, J., Gupta, A., Korenblum, D., Greenspan, H., Ma, Y., Rubin, D.: Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results. Radiology 256(1), 243–252 (2010)
4. Mougiakakou, S., Valavanis, I., Nikita, A., Nikita, K.: Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. Artificial Intelligence in Medicine 41, 25–37 (2007)

5. Duda, D., Kretowski, M., Bezy-Wendling, J.: Texture Characterization for Hepatic Tumor Recognition in Multiphase CT. Biocybernetics and Biomedical Engineering 26(4), 15–24 (2006)
6. Ye, J., Sun, Y., Wang, S.: Multi-Phase CT Image Based Hepatic Lesion Diagnosis by SVM. In: 2nd International Conference on Biomedical Engineering and Informatics, pp. 1–5 (2009)
7. Quatrehomme, A., Hoa, D., Subsol, G., Puech, W.: Review of Features Used in Recent Content-Based Radiology Image Retrieval Systems. In: Proceedings of the Third International Workshop on Image Analysis, pp. 105–113 (2010)
8. Deepa, S.N., Devi, B.A.: A survey on artificial intelligence approaches for medical image classification. Journal of Science and Technology 4(11), 1583–1595 (2011)
9. Unser, M.: Sum and Difference Histograms for Texture Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(1), 118–125 (1986)
10. Laws, K.I.: Textured Image Segmentation. PhD thesis, University of Southern California (January 1980)
11. Witten, I.H., Frank, E., Hall, M.A.: CHAPTER 10 Introduction to Weka. In: Data Mining, 3rd edn., pp. 403–406. Morgan Kaufmann (2011)
12. Platt, J.: Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In: Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)
13. Nino-Murcia, M., Olcott, E., Jeffrey, R.J., Lamm, R., Beaulieu, C., Jain, K.: Focal liver lesions: pattern-based classification scheme for enhancement at arterial phase CT. Radiology 215, 746–751 (2000)

VISCERAL: Towards Large Data in Medical Imaging — Challenges and Directions

Georg Langs[1], Allan Hanbury[2], Bjoern Menze[3], and Henning Müller[4]

[1] Computational Image Analysis and Radiology Lab Department of Radiology, Medical University of Vienna
[2] Institute for Software Technology and Interactive Systems Faculty of Computer Science, Vienna University of Technology
[3] Computer Vision Lab, ETH Zurich
[4] University of Applied Sciences Western Switzerland, Sierre (HES-SO)

Abstract. The increasing amount of medical imaging data acquired in clinical practice holds a tremendous body of diagnostically relevant information. Only a small portion of these data are accessible during clinical routine or research due to the complexity, richness, high dimensionality and size of the data. There is consensus in the community that leaps in this regard are hampered by the lack of large bodies of data shared across research groups and an associated definition of joint challenges on which development can focus. In this paper we describe the objectives of the project VISCERAL. It will provide the means to jump–start this process by providing access to unprecedented amounts of real world imaging data annotated through experts and by using a community effort to generate a large corpus of automatically generated standard annotations. To this end, VISCERAL will conduct two competitions that tackle large scale medical image data analysis in the fields of anatomy detection, and content–based image retrieval, in this case the retrieval of similar medical cases using visual data and textual radiology reports.

Keywords: Medical imaging, Large scale data, Localization, Retrieval.

1 Introduction

The amount and complexity of information present in medical imaging data on a hospital scale is enormous. Part of this information has immediate diagnostic relevance, part becomes relevant only when studied in the context of a large cohort (e.g., when studying subtle characteristics of diseases such as mild cognitive impairment), and part might not only be relevant at the time of acquisition, but also when used as reference during later radiological assessment. In the context of both *computer aided diagnosis (CAD)*, and medical research, a large research community has emerged that tackles the extraction and quantification of task specific relevant information from medical imaging data [1]. Traditional problems in this domain are the segmentation of anatomical structures, the detection of pathology, or the measurement of specific markers that correlate with disease or treatment outcome.

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 92–98, 2013.

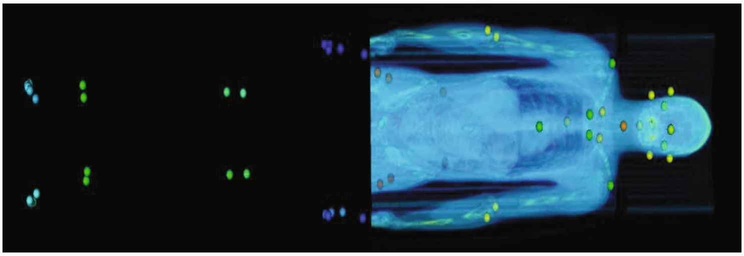


Fig. 1. Can we learn more than predefined targets from medical imaging data ?

The natural variability in these data and the corresponding difficulty in identifying anomalities that may be only subtle deviations from a healthy cohort results in a large body of work focusing on specific diseases. Typically approaches, such as CAD rely on a well controlled set of examples and corresponding expert annotations [2, 3].

This *paradigm* has several limitations. First, it typically focuses on replicating expert judgements learned during the training phase. More importantly, it does not scale to amounts of data necessary to understand diseases with subtle and complex effects that would require a much larger set of annotated examples in order to represent the difference between control– and patient group well. Both limitations together result in a severe hampering of efforts to identify novel markers, that are not part of the annotated training corpus or existing clinical routine.

1.1 Towards Large Data

Scaling to large data is necessary for computational approaches dealing with a wide range of diseases. It is a prerequisite to understand subtle characteristics of populations and to computationally identify novel markers. Among the challenges we face, and for which we lack adequate methodology, are:

- Making information in large data accessible during clinical routine (e.g., radiological assessment).
- Learning models of diseases and corresponding imaging markers with no or only limited supervision.
- Fast and efficient collection of large amounts of data relevant for a specific question.
- Leveraging the amount of data, and the relatively confined domain of the human body in an optimal way, when analyzing medical imaging data, and when identifying sublte markers, and relationships.
- Using both radiology report- and image information simultaneously, when searching for findings in images, which were part of clinical routine.

These are only a few of the challenges we face, and whose solution would be a crucial step towards harvesting the information present in medical images, that currently remains unused. To tackle these questions we need novel methodology.

A parallel development taking place in the computer vision community suggests that the step from small- to large data is accompanied by substantial changes in approaching tasks such as learning, representation, or retrieval [4]. It coincides with a shift from supervised learning methods to semi–supervised or even unsupervised training. An example is the aim to scale beyond hundreds of thousands of images and hundreds of categories in the context of the Internet. This has shown promising results in [5]. Unsupervised learning approaches for category models from large image data sets have been explored, too. Examples are [4, 6–8].

While information extraction from images in 2D is an active field of research by far the largest amount of visual data produced in hospitals are multi–slice series of images [2]. Benchmarks in medical imaging exist with ImageCLEF [9] but focusing rather on 2D and textual image retrieval, whereas the largest of data produced is currently 3D tomographic data. Medical imaging data are estimated to have reached 30% of the overall world wide data storage in 2010 [10].

In this paper we outline the main challenges we face when working with and analyzing large medical imaging data, and suggest two primary directions where advances are needed. We propose two corresponding challenges for discussion. They will be organized in the EU funded VISCERAL project to help focus the efforts in the community and to offer a means to compare methodology across research groups worldwide.

2 Open Questions in Medical Imaging and Directions Proposed

Recently, interest in alternatives and conceptual extensions to traditional CAD systems has emerged. Among others, an example of a particularly promising direction is the use of image retrieval that instead of providing a direct automated assessment allows for the efficient search for comparable and relevant cases. These cases are presented to aid the physician who is performing reading and interpretation of the case at hand. The visual content of medical images has been used for information retrieval for over 15 years [11] and has shown to improve quality of diagnosis [12]. Visual retrieval can extract rich information beyond the associated textual cues and is a promising direction to make use of the medical imaging databases in hospitals. Often, image retrieval is is mixed up with very simple tasks of classifying images into modality and anatomic region. Such a classification as preprocessing can not really be seen as a retrieval process and is rather the first step for the extraction of information usable to provide matching cases, anatomical regions or mine the data for specific pathologies.

Methodological challenges of interest include:

1. Scalability to large amounts of data. What is necessary to work on real data of a PACS and thus on very large and heterogeneous data sets, which have never been analyzed at such a large scale for retrieval of medical visual data as of yet? The VISCERAL project will extend the medical image analysis

towards very large data sets (hundreds of categories, and many thousands of data sets), which makes use of a new families of methods (unsupervised learning, modeling, and categorization) necessary. These methods have until now mainly been used in the computer vision community but only little for medical data.
2. Unsupervised and autonomous learning: the scale of the data (many TeraBytes) makes the fully autonomous and un–supervised building of models and categorization/indexing of the data crucial.
3. What is the right interface for injecting prior knowledge on anatomy and other structural elements into this algorithmic analysis?
4. Efficient annotation on pre–processed data instead of raw images to facilitate annotation. Making use of semi–supervised strategies instead of supervised learning on limited training bodies.
5. Generalization power to a large and diverse set of data, and the inclusion of a potentially growing set of training images during the learning phase.
6. Introducing image–based search and retrieval as an alternative to computational classification and traditional computer aided diagnosis that is concentrating on single specific phenomena.

3 Two Challenges to Focus Research Efforts

We propose two challenges, with the aim of spurning discussion, and refinement based on the response by the community. Challenges will be based on large amounts of partially annotated data.

3.1 Competition 1

Anatomical structure identification, detection and segmentation competition on a full body scale for fully-automated processing in the clinic. Participants are provided with 3D image data (multimodal full body scans, volumes containing specific anatomical structures as encountered in clinical practice) together with training annotations on a subset of the data. For evaluation, test data consisting of 3D volumes will be processed by the participants algorithms. The objective is to identify, localize and segment the anatomical structures present in the data. We will evaluate both with regard to comprehensive identification and to subset localization, in order to be able to include algorithms developed for specific organs as a secondary task within the competition

3.2 Competition 2

Similar case retrieval to allow tools and algorithms to be evaluated on real clinical tasks with a potentially larger impact. Given a query case consisting of either image, volume or potentially additional textual information the objective is to retrieve similar cases in terms of characteristics such as diagnosis or differential diagnosis. Challenges that have to be solved include incorrect or incomplete data (for example, data that was not entered into the record by a physician), and potentially very small regions relevant for the similarity computation (for example, a small fracture in an entire thorax CT).

3.3 Secondary Impact

Besides the immediate impact on research in the two suggested direction the VISCERAL project has the potential to advance the state of the art in several related questions:

- Allow medical image analysis on a very large scale (the scale of a PACS in many regional hospitals) and compare the results across many research groups;
- Create ground truth based on partial manual annotation and then also based on the results of the participating systems, also allowing to scale from purely manual annotation to mixed approaches of manual and semi-automatic ground truthing;
- Develop an infrastructure for medical image data sharing and potentially also sharing of components of participants in the cloud;
- Compare techniques for quality and speed across many research groups based on the same data and tasks, allowing to identify the most promising approaches to several current research challenges;
- Potentially allow for the creation of approvals for image retrieval as diagnosis aid if a reference database and a proof of quality can be shown.

4 Corpora

In parallel to serving as a benchmark to compare methods across a large number of possible alternatives, the challenges will serve as the basis for two corpora that collect imaging- and text data together with expert annotations, and semi-automated annotations achieved during the competitions. Part of the data will be annotated by experts, to obtain a *Gold Corpus.* At the same time, every participant will be contributor to a *Silver Corpus.* The latter will be formed by deriving a consensus across the entries of the participants.

The generation of the silver corpus will require methods for consolidating annotations obtained from independent groups participating in the competition. In the competition, algorithms are to be trained or tuned on the gold corpus train data; test data will be the silver corpus data and the gold corpus test data. Averaging the labels over the silver corpus data will lead to a fused label that is better than the individual label estimate, if estimates are unbiased. We will measure bias of the labels on the gold corpus test data that will be hidden in the test set. This will require a tool that also returns general quality measure (bias, variance) for individual image volumes, as well as for individual contributors of labels sets (i.e., algorithms).

5 Infrastructure

To allow for the distribution of very large data sets a new type of infrastructure seems necessary. 10 TB of data can not easily be downloaded or shipped on hard disk to many participants. Could computing on the other hand is often proposed for dealing with large storage and computing capacities. VISCERAL will make these capacities available to the medical imaging community and give participants in the benchmark access to a virtual machine (Windows on Linux) in the cloud to access the training data and prepare the systems. Then, the execution of the tools on the text data will be started by the organizers and results can subsequently be compared based on the runs of the test data. This allows to compare tools not only based on classification or retrieval accuracy but also based on speed or efficiency criteria, which becomes increasingly important when dealing with very large data sets. This should also give equal opportunities to groups from poorer and richer countries as all resources are controlled. The fact that only access to training data is given to the participants means that non–anonymous data sets could be distributed and used in the same way potentially. Bringing the algorithms to the data and not the data to the algorithms seems the only feasible approach when looking at extremely large–scale challenges.

6 Conclusion

Medical image analysis has brought many new interesting and successful techniques. Over the past 30 years it has helped to develop novel tools to aid diagnosis, that have a substantial impact on diagnosis quality, and treatment success. The quickly increasing amount of data produced and digitally available poses new challenges. How can we make use of these data and how can we exploit knowledge being stored in past data.

At the same time, quickly rising costs in health care will make it necessary to use all available data in the best possible way to take case of new patients. Past data can help in this process, keeping the privacy of patients protected at the same time.

In the VISCERAL project we propose two challenges to the medical imaging research community. Both challenges will use medical imaging data in a new scale, making in the order of 10 TB available for research. Both challenges concentrate on specific aspects in the image analysis chain. The first challenge concentrates on extracting anatomic regions and locations from the data, which is necessary in many contexts, such as for all further steps that compare tissue and abnormalities within the same anatomic regions of several patients. The second challenge focuses on the retrieval of similar cases: pathology–oriented retrieval. Algorithms that tackle the latter challenge can -although not traditional CAD - by viewed as a diagnosstic aid in fields such as evidence–based medicine where studies related to a specific patient need to be found and case–based reasoning where similar cases are compared to a patient being treated.

A crucial component of VISCERAL is the participation of the community at an early stage, when specifying and refining the tasks, and benchmakring measures, in order to truly support relevant research. and to allow VISCERAL to contribute to a leap forward in bringing more medical imaging to the clinical workflow.

References

1. Doi, K.: Current status and future potential of computer–aided diagnosis in medical imaging. British Journal of Radiology 78, 3–19 (2005)
2. Depeursinge, A., Vargas, A., Platon, A., Geissbuhler, A., Poletti, P.–A., Müller, H.: 3D Case–Based Retrieval for Interstitial Lung Diseases. In: Caputo, B., Müller, H., Syeda-Mahmood, T., Duncan, J.S., Wang, F., Kalpathy-Cramer, J. (eds.) MCBR-CDS 2009. LNCS, vol. 5853, pp. 39–48. Springer, Heidelberg (2010)
3. Chen, W., Giger, M.L., Li, H., Bick, U., Newstead, G.M.: Volumetric texture analysis of breast lesions on contrast–enhanced magnetic resonance images. Magnetic Resonance in Medicine 58(3), 562–571 (2007)
4. Torralba, A., Fergus, R., Freeman, W.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11), 1958–1970 (2008)
5. Oliva, A., Torralba, A.: Building the Gist of a Scene: The Role of Global Image Features in Recognition. Visual Perception (2006)
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
7. Langs, G., Donner, R., Peloschek, P., Bischof, H.: Robust Autonomous Model Learning from 2D and 3D Data Sets. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 968–976. Springer, Heidelberg (2007)
8. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
9. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., Seco de Herrera, A.G., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum) (September 2011)
10. Riding the wave how europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data. A submission to the European Comission (October 2010)
11. Lowe, H.J., Antipov, I., Hersh, W., Smith, C.A.: Towards knowledge–based retrieval of medical images. The role of semantic indexing, image content representation and knowledge–based retrieval. In: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Nashville, TN, USA, pp. 882–886 (October 1998)
12. Aisen, A.M., Broderick, L.S., Winer-Muram, H., Brodley, C.E., Kak, A.C., Pavlopoulou, C., Dy, J., Shyu, C.R., Marchiori, A.: Automated storage and retrieval of thin–section CT images to assist diagnosis. System Description and Preliminary Assessment 228(1), 265–270 (2003)

Customised Frequency Pre-filtering in a Local Binary Pattern-Based Classification of Gastrointestinal Images

Sebastian Hegenbart[1], Stefan Maimone[1], Andreas Uhl[1],
Andreas Vécsei[2], and Georg Wimmer[1]

[1] Department of Computer Sciences
University of Salzburg
Salzburg, Austria
[2] St. Anna Children's Hospital
Vienna, Austria
{gwimmmer,uhl}@cosy.sbg.ac.at

Abstract. Local Binary Patterns (LBP) is a widely used approach for medical image analysis. Limitations of the LBP operator are its sensitivity to noise and its boundedness to first derivative information. These limitations are usually balanced by extensions of the classical LBP operator (e.g. the Local Ternary Pattern operator (LTP) or the Extended LBP (ELBP) operator). In this paper we present a generic framework that is able to overcome this limitations by frequency filtering the images as pre-processing stage to the classical LBP. The advantage of this approach is its easier adaption and optimization to different application scenarios and data sets as compared to other LBP variants. Experiments are carried out employing two endoscopic data sets, the first from the duodenum used for diagnosis of celiac disease, the second from the colon used for polyp malignity assessment. It turned out that high pass filtering combined with LBP outperforms classical LBP and most of its extensions, whereas low pass filtering effects the results only to a small extent.

Keywords: LBP, frequency filtering, medical image processing, endoscopy.

1 Introduction

Computer-aided decision support systems relying on automated analysis of medical imagery receive increasing attention [1]. Among other techniques, feature extraction employing local binary patterns (LBP) is a popular approach which has been used for a wide variety of medical application domains, including colon polyp detection and classification [2], the diagnosis of celiac disease [3], and detection of gastric cancer [4] in classical flexible endoscopy, detection of ulcers [5], tumors [6], and blood [7] in capsule endoscopy images, and breast cancer diagnosis in microscopic specimens [8]. Even in traditional Chinese medicine (TCM) LBP have been applied, to distinguish between cold gastritis and heat gastritis in gastroscopic images [9].

The classical LBP operator has some limitations, as it is (i) sensitive to noise and (ii) can only reflect first derivative information since it is directly derived from the image intensity function. With respect to the first issue (i), the Local Ternary Pattern operator

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 99–109, 2013.

(LTP) [10] has been introduced which uses a thresholding mechanism which implicitly improves the robustness against noise (a sort of implicit denoising is applied restricting the attention to low-frequency information), being especially important for many medical imaging modalities. Closely related are techniques applying the LBP operator to the approximation subband of a wavelet transform [3,11], since these data represent low-pass information as well which is hardly affected by noise. The second issue (ii) is tackled by introducing the Extended LBP (ELBP) [12] which applies a Gradient filter to the image before deriving LBP histograms. Closely related are techniques applying the LBP operator to detail subbands of a wavelet transform [13,14]. Some techniques attempt to deal with both issues concurrently, e.g. the Extended Local Ternary Pattern operator (ELTP [3]) integrates ELBP and LTP, and wavelet techniques including LBP information derived from approximation and details subbands (WTLBP), respectively [3,11].

All these proposed LBP extensions share the problem that their usage is somehow adhoc and that they are highly non-trivial to optimize to some specific setting, especially with respect to the extent of the frequency band the attention is restricted to. For example, classical wavelet subbands allow only an octave-based partitioning of frequency space and the usage of a spatial domain gradient operator does not offer siginificant adaptation potential at all.

These observations motivate the approach followed in this paper. As a generic framework we introduce a pre-filtering stage, in which we apply Fourier-based high-pass and low-pass filtering, respectively, before the LBP operator is applied to the filtered image material. In this manner, we have absolute control over the chosen cut-off frequency of the corresponding filter and are perfectly able to customize the frequency band the LBP operator is being applied to. We assess the effect of varying cut-off frequency and compare results of our generic framework to the abovementioned specialised LBP variants. Frequency filtering and the LBP operator are widely used applications for image processing, but the proposed combination of these two applications has not been proposed so far.

This paper is structured as follows. Section 2 explains the Fourier-domain filtering, as well as the LBP approach. Experimental results are shown in Section 3, where we provide comparatative classification results employing two endoscopic data sets, the first from the duodenum used for diagnosis of celiac disease, the second from the colon used for polyp malignity assessment. Section 4 concludes the paper and provides outlook to further refine the proposed approach.

2 Feature Extraction

2.1 Image Filtering in the Frequency Domain

In this section we shortly review image filtering to consider only their respective high frequency or low frequency information. Filtering is applied in the Fourier frequency domain.

A low pass filter in the frequency domain intuitively means zeroing all frequency components above a cut-off frequency and a high pass filter means zeroing all frequency components beneath a cut-off frequency. We are normalizing the images frequencies so that they are between zero and one. That means for example that a low pass filter with cutoff frequency 1 is zeroing nothing and a low pass filter with cutoff frequency 0 is zeroing all frequencies. An ideal low pass filter (ILPF) is defined as follows:

$$h(u,v) = \begin{cases} 1 & \text{if } D(u,v) < D_0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where D_0 is the cut-off frequency and

$$D(u,v) = \frac{\sqrt{(u - M/2)^2 + (v - N/2)^2}}{\sqrt{(M/2)^2 + N/2)^2)}} \tag{2}$$

is the normalized distance of (u,v) ($u \in \{0,\ldots,M\}$, $v \in \{0,\ldots,N\}$) to the center of the $(M+1) \times (N+1)$ image $I(u,v)$. The ideal high pass filter (IHPF) is defined conversly to the ILPF. Using ILPF and IHPF can result in unwanted ringing which is caused by the sharp transition from stop to pass band. To avoid this effect, one can smooth the transition between the stop and pass band. One example for a smoother filter is the Butterworth filter [15]. The Butterworth low pass (BLPF $h(u,v)$) and high pass filters (BHPF $g(u,v)$) are defined as follows:

$$h(u,v) = \tfrac{1}{1+(D(u,v)/D_0)^{2n}}, \qquad g(u,v) = \frac{1}{1+(D_0/D(u,v))^{2n}}. \tag{3}$$

The higher the order n, the sharper the transition between the stop and pass band. For $n \to \infty$, the Butterworth LP (HP) filter converges toward the ILPF (IHPF).

In Figure 1 we see examples of filtered images. As cut-off frequencies we employ 0.06 for the high pass filters and 0.1 for the low pass filters.

Especially in case of the low pass filtered images we can see that the ILPF causes ringing while the BLPF avoids this effect. The high pass filtered images of the boy are looking quite different to the orginal image of the boy, while the high pass filtered endoscopic images of a healthy duodenum are looking quite similar to the original endoscopic image. That is because the endoscopic image does not contain visually significant low pass information in the filtered band.

After the filtering process, the frequency filtered image

$$I_h(u,v) = I(u,v)h(u,v) \text{ or } I_g(u,v) = I(u,v)g(u,v) \tag{4}$$

is transformed back to the spatial domain to apply the LBP operator to it like described in the following section.

2.2 Local Binary Patterns

The basic Local Binary Patterns (LBP) operator was introduced to the community by Ojala et al. [16]. The LBP operator considers each pixel in a neighborhood separately. Hence the LBP could be considered a micro-texton. The operator is used to model a pixel neighborhood in terms of pixel intensity differences. This means that several common

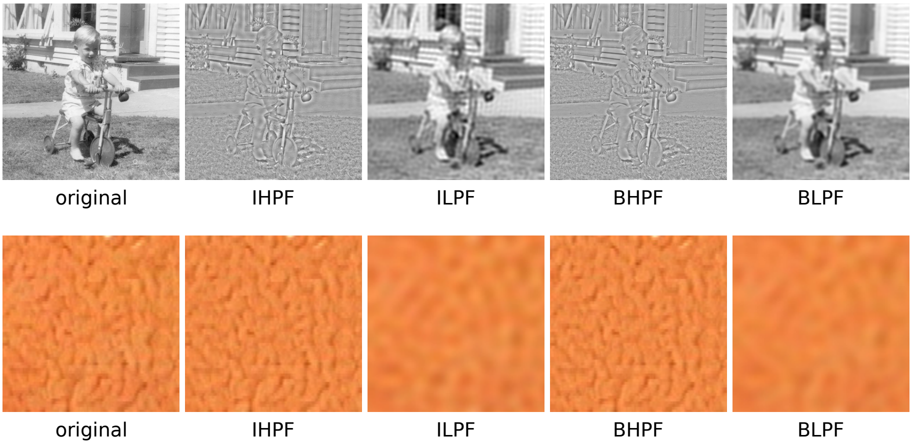


Fig. 1. Examples of filtered images, a common image and an example from the celiac-dataset (see below)

$$\begin{array}{|c|c|c|}\hline 61 & 73 & 81 \\ \hline 51 & 44 & 39 \\ \hline 42 & 57 & 71 \\ \hline \end{array} \xrightarrow{\text{neighborhood comparisons}} \begin{array}{|c|c|c|}\hline 1 & 1 & 1 \\ \hline 1 & & 0 \\ \hline 0 & 1 & 1 \\ \hline \end{array} \times \begin{array}{|c|c|c|}\hline 1 & 2 & 4 \\ \hline 128 & & 8 \\ \hline 64 & 32 & 16 \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline 1 & 2 & 4 \\ \hline 128 & & 0 \\ \hline 0 & 32 & 16 \\ \hline \end{array} \rightarrow \sum = 183$$

Fig. 2. Demonstration of calculating the LBP operator

structures within a texture are represented by a binary label. The joint distributions of these labels are then used to characterize a texture. The LBP operator is defined as

$$LBP_{r,p}(x,y) = \sum_{k=0}^{p-1} 2^k \, s(I_k - I_c). \tag{5}$$

I_k is the value of neighbour number k and I_c is the value of the corresponding center pixel. The s function acts as sign function, mapping to 1 if the difference is smaller or equal to 0 and mapping to 0 else. In Figure 2 we can see an example for the calculation of the basic LBP operator for $p = 8$ and $r = 1$.

The LBP histogram of an image I is formally defined as

$$H_I(i) = \sum_{x,y} (LBP_{r,p}(x,y) = i) \qquad i = 0, \cdots, 2^p - 1. \tag{6}$$

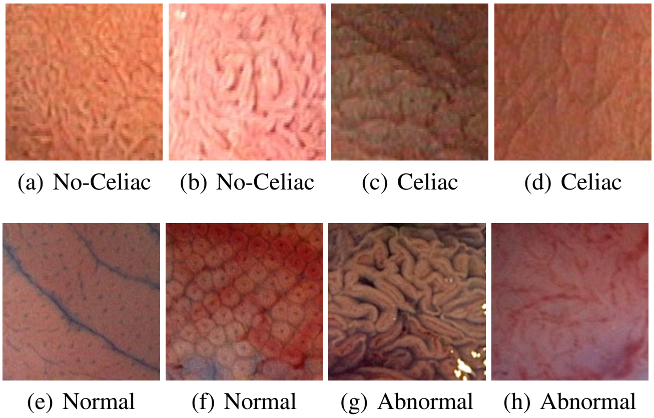
The basic operator uses an eight-neighborhood ($p = 8$) with a 1-pixel radius ($r = 1$). To overcome this limitation, we use the notion of scale as discussed by Ojala et al. [17] by applying averaging filters to the image data before the operators are applied. Thus, information about neighboring pixels is implicitly encoded by the operator. The appropriate filter sizes for a certain scale is calculated as described by Mäenpää [18]. We use three scale levels for the original LBP operator and all LBP variants (as mentioned in the Introduction) throughout this work. The histograms of the three scale levels are concatenated to form the feature vector of an image. For more details about the employed LBP variants see e.g. [3].

3 Experimental Study

3.1 Experimental Setup

Experiments are carried out using two medical image databases. The images of both databases are classified by the k-NN classifier using histogram intersection as metric. In case that experiments are applied to color images, LBP histograms are generated for each color channel and are finally concatenated.

The Celiac Disease Image Database. Celiac disease is a complex autoimmune disorder in genetically predisposed individuals of all age groups after introduction of gluten containing food. The celiac state of the duodenum is usually determined by visual inspection during the endoscopy session followed by a biopsy of suspicious areas. Images used are of size 128×128 and are divided into two classes according to their histologic state. The class denoted as "No-Celiac" represents a healthy duodenum with normal crypts and villi, whereas the class denoted as "Celiac" represents a duodenum with mild or marked atrophy of the villi or the villi are even entirely absent (see Figure 3). Especially feature extraction methods using high frequency information could be interesting to differentiate between the two classes, since images of patients with celiac disease have less or entirely no villi and so a lower amount of contrast compared to images of patients without celiac disease. That is one of the motivations to employ frequency filtering (especially high pass filtering) as preprocessing step to the feature extraction by means of LBP.


(a) No-Celiac (b) No-Celiac (c) Celiac (d) Celiac

(e) Normal (f) Normal (g) Abnormal (h) Abnormal

Fig. 3. Images of the two classes from the duodenum ((a) – (d)) and from the colon ((e) – (h))

The results of classifying the celiac disease database [19] are computed using an evaluation and a training set to avoid overfitting (Table 1 lists the number of image samples and patients per class). An image of the evaluation set is classified to the class, where the majority of its k nearest neighbors from the training set belong to. The k for the k-NN classifier, used to classify the test set, is between 1 and 25 and is optimized in the training set (the k with the highest accuracy using leave–one–out cross–validation (LOOCV) on the training set is used).

The Polyp Image Database. Polyps of the colon are a frequent finding and are usually divided into metaplastic, adenomatous and malignant [2]. A high magnifying colonoscope (magnification factor 150) is used to obtain images of the polyp's surface under indigo carmine staining, since images have to be as detailed as possible to uncover the fine surface structure of the mucosa as well as small lesions. The class denoted as "Normal" represents normal colon mucosa or hyperplastic polyps (non-neoplastic lesions), the class denoted as "Abnormal" represents neoplastic, adenomatous and carcinomatous structures (see Figure 3). As we notice from Figure 3, the pits of images from class Normal are regular and tightly distributed and are shaped similarly, in contrast to those of class Abnormal. Also for the polyp image database feature extraction methods using high frequency information could be helpful to distinguish the two classes, since the edge infornmation is important to detect the pits.

All images are of size 256×256, Table 1 lists the number of image samples and patients per class.

Table 1. Number of image samples per class of the two image databases (ground truth based on histology)

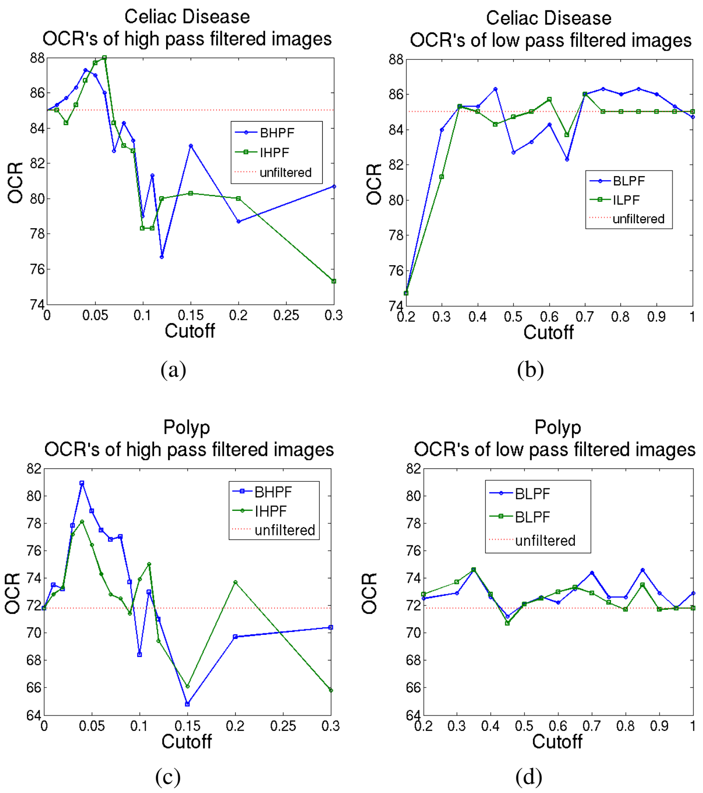
		Celiac Disease				
	Training set			Evaluation set		
Class	No-Celiac	Celiac	Total	No-Celiac	Celiac	Total
Number of images	155	157	312	151	149	300
Number of patients	66	21	87	65	19	84
		Polyp				
Class	Normal		Abnormal		Total	
Number of images	199		518		716	

In contrast to the celiac disease image database, the polyp image database [2] consists of images from too few patients (40) to devide them into an evaluation and a training set. To avoid overfitting as far as possible, we employ leave–one–patient–out cross–validation (LOPO) [19]. As k for the k-NN classifier, we use the one k between 1 and 25 with the highest overall classification rate (OCR).

3.2 Results

In Figure 4 we see four diagrams, which show how the cutoff frequency influences the results in terms of OCR accuracy, when frequency filtering is applied to the images of the two image databases followed by computing their LBP histograms. The red dotted line shows the result for the unfiltered images. The used Butterworth filters have order $n = 4$.

As we can see, the results of the high pass filtered images are superior to those of unfiltered images for lower cutoff frequencies. For higher cutoff frequencies, results are worse than the results of the unfiltered image. The unfiltered images and the low



(a) (b)

(c) (d)

Fig. 4. OCR accuracies of different cutoff frequencies for high pass and low pass filtered images of the two image databases

pass filtered images have similar results, except for cutoff's ≤ 0.3 in case of the celiac disease image database.

In Table 2 we show comparatative results (i.e. LBP with pre-filtering vs. LBP variants) for the celiac disease database. Only the highest results across all tested cutoff frequencies are shown.

LBP applied to high pass filtered images works quite well compared to other LBP-based methods. Especially in case of the grayscale images, LBP applied to IHP filtered images works better than any other LBP-based method apart from WTLBP. Also for color images, high pass filtered images provide good results, but here color does not improve the grayscale results indicating a feature dimensionality problem.

Table 3 shows the results for the polyp image database. Again, applying LBP to high pass filtered images provide better results than applying LBP to unfiltered or low pass filtered images. Especially the Butterworth variant is clearly superior in both the grayscale and RGB cases, respectively. Similar to the celiac disease dataset, the LBP & BHPF results are beaten only by LBP variants combining high pass and low pass filtering to some extent (i.e. ELTP, WTLBP).

Table 2. Results of the LBP-variants applied to original and frequency filtered (LBP & ...) images of the celiac disease image database

Colorspace		Grayscale			RGB	
Method	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
LBP & IHPF	91.3	84.8	88.0	89.3	79.5	84.3
LBP & BHPF	92.0	82.8	87.3	89.3	82.8	86.0
LBP & ILPF	89.9	82.1	86.0	85.2	80.8	83.0
LBP & BLPF	88.6	84.1	86.3	83.9	82.8	83.3
LBP	90.6	79.5	85.0	87.3	79.5	83.3
LTP	83.2	75.5	79.3	94.0	75.5	84.7
ELBP	94.0	74.2	84.0	93.3	81.5	87.3
ELTP	92.0	73.2	83.0	88.6	82.8	85.7
WTLBP	92.6	85.4	89.0	88.6	88.1	88.3

Table 3. Results of the LBP-variants applied to the original and frequency filtered (LBP & ...) images of the polyp image database

Colorspace		Grayscale			RGB	
Method	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
LBP & IHPF	96.9	28.8	78.1	96.0	43.3	84.1
LBP & BHPF	98.7	34.3	80.9	98.5	60.1	87.9
LBP & ILPF	96.0	18.7	74.6	96.1	59.1	85.9
LBP & BLPF	95.8	19.2	74.6	95.0	62.1	85.9
LBP	93.2	15.7	71.8	93.1	58.6	83.5
LTP	94.4	17.2	73.0	89.6	44.4	77.1
ELBP	92.5	37.9	77.4	96.1	56.7	85.2
ELTP	93.4	60.6	84.4	97.7	69.7	89.9
WTLBP	89.2	59.1	80.9	92.9	58.6	83.4

Finally, we want to assess statistical significance of our results. The aim is to analyze if the images from the two databases are classified differently or similarly by the various LBP-based methods. We use the McNemar test [20], to test if two methods are significantly different for a given level of significance α by building test statistics from correctly and incorrectly classified images, respectively. Tests were carried out with a significance level of $\alpha = 0.01$. It turned out that there are no significant differences in case of the celiac disease database, contrary to the polyp database, where significant differences occur. The results for the polyp database (RGB color space) are displayed in Figure 5. We can observe, that LBP & BHPF is significant different to LBP, contrary to the methods LBP and LBP & IHPF. Also LBP & BHPF is significant different to LBP & IHPF, indicating that smoothing the transition between the stop and pass band (to avoiding ringing effects) has (a positive) impact on the classification results of the images when using the LBP operator as feature extraction method. The methods ELTP and LTP are significantly different to other methods, whereat ELTP works better and LTP works worse than the other methods.

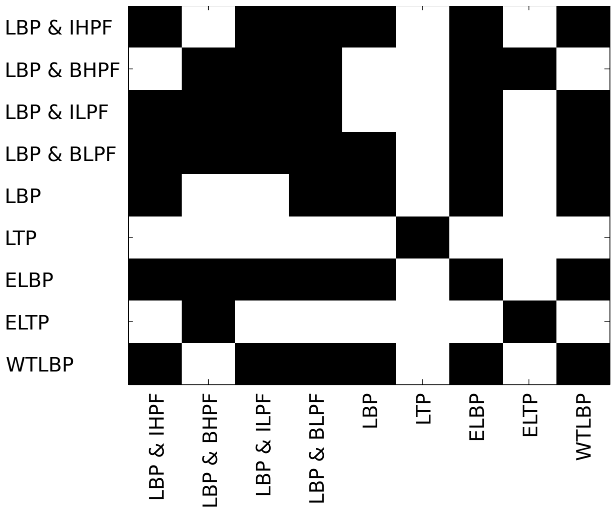


Fig. 5. Results of the McNemar test for the polyp database. A white square in the i'th row and j'th column or in the j'th row and i'th column of a plot means that the i'th and the j'th method are significantly different with significance level $\alpha = 0.01$. If the square is black than there is no significant difference between the methods.

4 Conclusion

We have found that the generic framework of pre-filtering images before applying LBP feature extraction provides good results for our two endoscopic databases. Especially using high pass filters outperforms classical LBP and provides results competitive to several LBP variants that have been proposed in literature. As we can see in [3], these LBP variants again outperform other classical medical image processing methods like Gabor Wavelets, Discrete Wavelet Transform or Dual-Tree Complex Wavelet Transform. So our proposed approach using high pass filters also works quite well compared to non-LBP based methods. An additional advantage of our approach is that it can be easily customised to a specific dataset by adjusting cut-off frequencies appropriately.

The reason we are focusing on high frequency information is, that high frequency information clearly points out the visible differences between healty and unhealthy mucosa on the two endoscopic data sets:

- In case of the celiac disease image database, images of patients with celiac disease have less or entirely no villi and so a lower amount of contrast compared to images of patients without celiac disease.
- In case of the polyp image database the edge information is important to detect the pits.

This is most likely the reason why our approach outperforms the classic LBP operator.

The proposed approach is only beaten by LBP variants which include both high pass and low pass filtering (WTLBP and ELTP) – based on this observation, we will consider in future work to use both high and low pass filtering in an adaptive manner where the optimal cut-off frequency is determined for both schemes (as indicated in Fig. 4) and the resulting LBP histograms are concatenated. Feature dimensionality can be controlled by applying feature subset selection techniques to bins of the resulting concatenated LBP histograms.

References

1. Liedlgruber, M., Uhl, A.: Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. IEEE Reviews in Biomedical Engineering (2011) (in press)
2. Häfner, M., Liedlgruber, M., Uhl, A., Vécsei, A., Wrba, F.: Color treatment in endoscopic image classification using multi-scale local color vector patterns. Medical Image Analysis 16(1), 75–86 (2012)
3. Vécsei, A., Amann, G., Hegenbart, S., Liedlgruber, M., Uhl, A.: Automated marsh-like classification of celiac disease in children using an optimized local texture operator. Computers in Biology and Medicine 41(6), 313–325 (2011)
4. Sousa, A., Dinis-Ribeiro, M., Areia, M., Coimbra, M.: Identifying cancer regions in vital-stained magnification endoscopy images using adapted color histograms. In: Proceedings of the 16th International Conference on Image Processing (ICIP 2009), Cairo, Egypt, pp. 681–684 (November 2009)
5. Li, B., Meng, M.Q.: Texture analysis for ulcer detection in capsule endoscopy images. Image and Vision Computing 27(9), 1336–1342 (2009)
6. Li, B., Meng, M.Q.H.: Small bowel tumor detection for wireless capsule endoscopy images using textural features and support vector machine. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, USA, pp. 498–503 (October 2009)
7. Li, B., Meng, M.Q.: Computer aided detection of bleeding regions for capsule endoscopy images. IEEE Transactions on Bio-Medical Engineering 56(4), 1032–1039 (2009)
8. Zhang, B.: Breast cancer diagnosis from biopsy images by serial fusion of random subspace ensembles. In: 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI), vol. 1, pp. 180–186 (October 2011)
9. Xu, Z., Guo, H., Chen, W.: Gastritis cold or heat image research based on lbp. In: 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE), vol. 5, pp. 331–333 (August 2010)
10. Tan, X., Triggs, B.: Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. In: Zhou, S.K., Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 168–182. Springer, Heidelberg (2007)
11. Wang, Y., Chun Mu, Z., Zeng, H.: Block-based and multi-resolution methods for ear recognition using wavelet transform and uniform local binary patterns, pp. 1–4 (December 2008)
12. Huang, X., Li, S., Wang, Y.: Shape localization based on statistical method using extended local binary pattern. In: Proceedings of the 3rd International Conference on Image and Graphics (ICIG 2004), Hong Kong, China, pp. 1–4 (2004)
13. Liu, X., You, X., Cheung, Y.: Texture image retrieval using non-separable wavelets and local binary patterns. In: International Conference on Computional Intelligence and Security, CIS 2009, pp. 287–291. IEEE Computer Society (2009)
14. Su, Y., Tao, D., Li, X., Gao, X.: Texture representation in aam using gabor wavelet and local binary patterns. In: SMC 2009: Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, pp. 3274–3279. IEEE Press, Piscataway (2009)
15. Butterworth, S.: On the theory of filter amplifiers. Wireless Engineer 7, 536–541 (1930)

16. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29(1), 51–59 (1996)
17. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution Gray-Scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
18. Mäenpää, T.: The local binary pattern approach to texture analysis - extensions and applications. PhD thesis, University of Oulu (2003)
19. Hegenbart, S., Uhl, A., Vécsei, A.: Systematic Assessment of Performance Prediction Techniques in Medical Image Classification A Case Study on Celiac Disease. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 498–509. Springer, Heidelberg (2011)
20. McNemar, Q.: Note on the sampling error of the difference between correlated proportions of percentages. Psychometrika 12(2), 153–157 (1947)

Bag–of–Colors for Biomedical Document Image Classification

Alba García Seco de Herrera, Dimitrios Markonis,
and Henning Müller

University of Applied Sciences Western Switzerland (HES–SO),
Sierre, Switzerland
{alba.garcia,dimitrios.markonis,henning.mueller}@hevs.ch
http://medgift.hevs.ch

Abstract. The number of biomedical publications has increased noticeably in the last 30 years. Clinicians and medical researchers regularly have unmet information needs but require more time for searching than is usually available to find publications relevant to a clinical situation. The techniques described in this article are used to classify images from the biomedical open access literature into categories, which can potentially reduce the search time. Only the visual information of the images is used to classify images based on a benchmark database of ImageCLEF 2011 created for the task of image classification and image retrieval. We evaluate particularly the importance of color in addition to the frequently used texture and grey level features.

Results show that bags–of–colors in combination with the Scale Invariant Feature Transform (SIFT) provide an image representation allowing to improve the classification quality. Accuracy improved from 69.75% of the best system in ImageCLEF 2011 using visual information, only, to 72.5% of the system described in this paper. The results highlight the importance of color for the classification of biomedical images.

Keywords: bag–of–colors, SIFT, image categorization, ImageCLEF.

1 Introduction

The number of biomedical articles published grew at a double–exponential pace between 1986 and 2006 according to [1]. Images represent an important part of the content in many publications and searching for medical images has become common in applications such as Goldminer[1], particularly for radiologists. Image retrieval has shown to be complementary to text retrieval approaches and images can well help to represent the content of scientific articles, particularly in applications using small interfaces such as mobile phones [2].

Many physicians have regular information needs during clinical work, teaching preparation and research activities [3,4]. Studies showed that the time for

[1] http://goldminer.arrs.org/

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 110–121, 2013.

answering an information need with MedLine is around 30 minutes [5], while clinicians state to have approximately five minutes available [6]. Finding relevant information quicker is thus an important task to bring search into clinical routine. To facilitate searching for images in biomedical articles, search engines such as Goldminer include the ability to filter search results by modality, age group or gender [7]. Imaging modalities can include typical classes such as x–ray, computed tomography (CT) or magnetic resonance imaging (MRI). In the biomedical literature, other classes such as photos (e.g. photomicrographs and endoscopic images), graphics (e.g. charts and illustrations) and compound figures also occur frequently [7,8,9]. For the modality classification, caption information can help if captions are well controlled like in the radiology domain but the more general biomedical literature makes it hard to find the modality information in the caption. Past work has shown that the image modality can be extracted from the image itself using visual features, only [10,11,12]. Therefore, in this paper, purely visual methods are used for the classification.

Our focus is on implementing, evaluating and developing visual features for representing images for the task of modality classification. To classify images many features based on color [13], texture [14], or shape [15] have been used. Although color information is important many approaches use only grey level information such as the Scale Invariant Feature Transform (SIFT) [16]. Additionally, several color image descriptors have been proposed [17]. Recently, a color extension to the SIFT descriptor was presented by van de Sande et al. [18]. Ai et al. [19] also proposed a color independent components based SIFT descriptor (CIC–SIFT) for image classification. Color and geometrical features combined are expected to improve the results for classification.

This paper extends image classification with SIFT features [20] by adding color features using bags–of–colors (BoC) based on [21] to represent the images. SIFT showed to be one of the most robust local feature descriptors with respect to geometrical changes [22]. As it contains only grey level information we fused results with BoC to include both color and texture information. The ImageCLEF 2011 database for medical modality classification was used as results for many research groups using state–of–the–art techniques on this database are available as baselines [23]. Both visual and textual approaches are possible and this paper concentrates on purely visual approaches.

The rest of the paper is organised as follows: Section 2 provides a detailed description of the methods and tools used. Section 3 reports on results while Section 4 concludes this work.

2 Methods

This section describes the dataset and the evaluation methodology used in this article. The main techniques and tools used are also detailed.

2.1 Dataset

The database of the medical ImageCLEF 2011 task [2] [23] is used in this document. The entire database consists of over 230,000 images of the biomedical open access literature. For modality classification 1,000 training and 1,000 test images were made available with class labels and a standard setup. Labels are one of 18 categories including graphs and several radiology modalities (see Figure 1). The sample images presented in Figure 2 demonstrate the visual diversity of the classes of the data set. Images are unevenly distributed across classes, which can affect the training of the classifiers and the resulting performance. In our study, a subset of 100 training images uniformly distributed across the classes was used for the creation of the visual vocabularies.

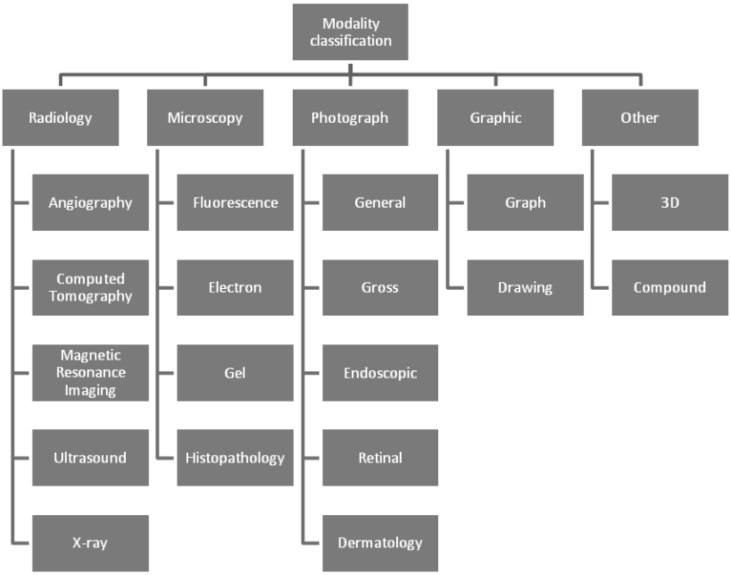


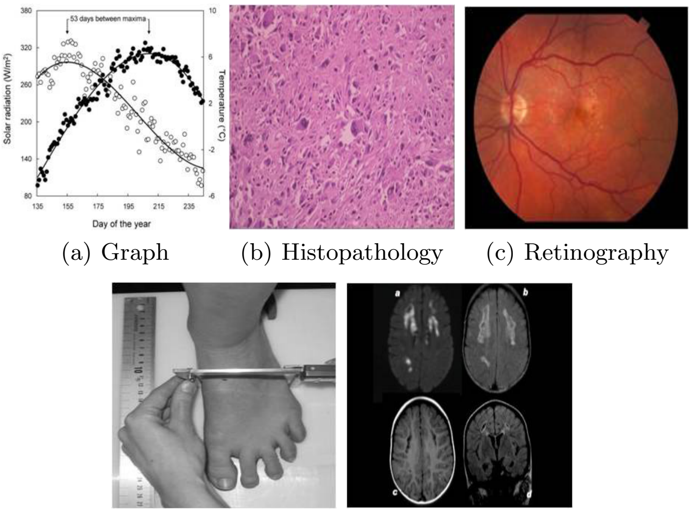
Fig. 1. Modality categories of the ImageCLEF 2011 medical task

2.2 The CIELab Color Space

We used the CIE (International Commission on Illumination) 1976 L*a*b (CIELab) space for our method because it is a perceptually uniform color space recommended by CIE[3] and used in many applications [24]. CIELab is a 3–D component space defined by L for luminance and a, b for the color–opponent dimensions for chrominance [24,25].

[2] http://imageclef.org/2011/medical/

[3] CIE is the primary organization responsible for standardization of color metrics.



(a) Graph (b) Histopathology (c) Retinography

(d) General Photo (e) Radiology

Fig. 2. Sample images from ImageCLEF 2011 medical data set including their class labels

2.3 Bags–of–Colors

Bags–of–Colors (BoC) is a method to extract a color signature from images introduced by [21]. The method is based on the Bag–of–Visual–Words (BoVW) image representation [26]. Each image is represented by a BoC from a color vocabulary C previously learned on a subset of the collection.

A color vocabulary $C = \{c_1, \ldots, c_{k_c}\}$, with $c_i = (L_i, a_i, b_i) \in CIELab$ is constructed by first finding the most frequently occurring colors in each image of the subset of the collection. In our case, frequent colors of the 100 selected images are used. A color is considered frequent if it occurs more than once for every 10,000 pixels in an image.The selected colors are clustered using a k–means algorithm [27]. We use for our experiments mainly $k_c = 200$ found by an analysis on the training set (see Table 1). For illustrations (Figures 4 and 5) the visual vocabularies, so the cluster centers of the color clusters for $k_c = 10, 20$ are shown including the histograms for example image types.

The BoC of an image I is defined as a vector $h_{BoC} = \{\bar{c}_1, \ldots, \bar{c}_k\}$ such that, for each pixel $p_k \in I \ \forall k \in \{1, \ldots, n_p\}$, with n_p being the number of pixels of the image I:

$$\bar{c}_i = \sum_{k=1}^{n_p} \sum_{j=1}^{n_p} g_j(p_k) \ \forall i \in \{1, \ldots, k_c\}$$

where

$$g_j(p) = \begin{cases} 1 \ if \ d(p, c_j) \leq d(p, c_l) \ \forall l \in \{1, \ldots, k_c\} \\ 0 \ otherwise \end{cases} \tag{1}$$

and $d(x, y)$ being the Euclidean distance between x and y.

Generally speaking, given a color vocabulary $C = \{c_1, \ldots, c_{k_c}\}$ defined by automatically clustered color occurrences in the CIELab color space, a BoC of an image is obtained by finding for each pixel of the image the closest color in the color vocabulary. The number of times each color appears in the image is then entered into a color histogram. The procedure is the following:

1. Convert images into the CIELab color space.
2. Create a color vocabulary:
   2.1. Find frequently occurring colors in each image from the 100 images selected.
   2.2. Cluster colors using the k–means algorithm.
3. Create a BoC for each image:
   3.1. Select for each pixel of each image, the closest color in the vocabulary using the Euclidean distance.
   3.2. Increment the corresponding bin of the output k_c–dimensional histogram.
4. Normalize to make the vectors comparable.

This procedure is described graphically in Figure 3.

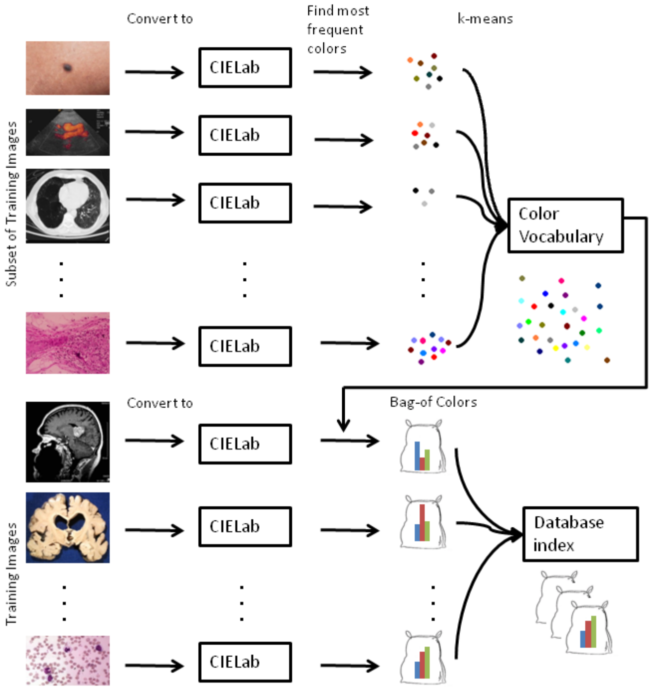


Fig. 3. The procedure for constructing the BoC

2.4 SIFT

SIFT [16] has been used for general image retrieval and also for medical image classification by several authors in the past [20]. The Fiji image processing package[4] was used for the extraction of the SIFT features. In this work, we use SIFT features represented as BoVW. For the creation of the vocabulary, our implementation of the DENCLUE (Density Clustering) [28] algorithm was used to increase the speed of the clustering.

2.5 Representation and Fusion

The images are represented as histograms and the similarity between images is calculated by comparing their histograms. The distance measure used in this article is the histogram intersection [29].

Late fusion [30] is used to combine the results of SIFT and BoC. First we obtain similarity scores separately using SIFT BoVW and BoC descriptors. Then, these scores are fused by the score–based CombMNZ fusion rule [31]. The image is classified into one class by a weighted k–NN voting [32]. We show results with varying k_c and k_{nn} in Tables 1 and 2.

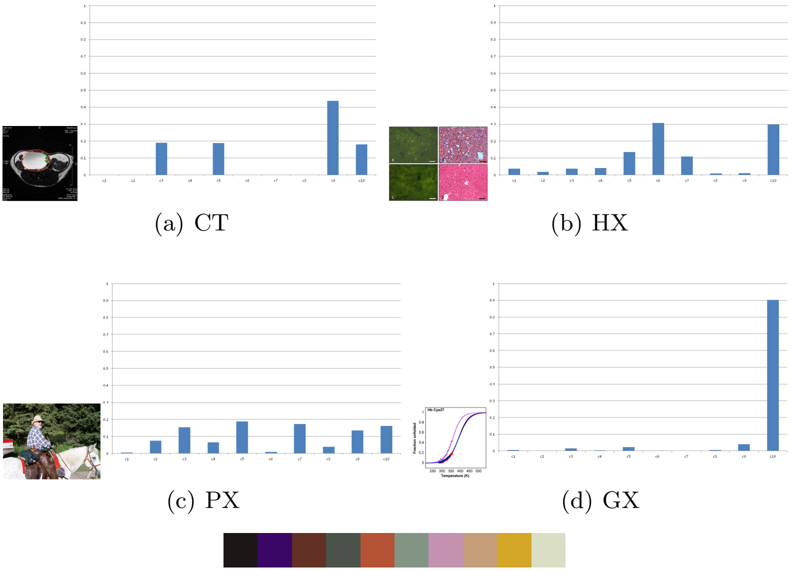
As an evaluation measure, accuracy was used, giving the percentage of correctly classified images over the entire test set of 1,000 images. This procedure allows for a fair comparison of the different schemes with and without the use of BoC.

3 Results

To analyze results using the BoC, we present two examples with 10 and 20 color terms (see Figures 4 and 5).

Given $k_c = 10$, a vocabulary $C_{10} = \{c_1, \ldots, c_{10}\}$ is created. The vocabulary contains ten colors corresponding to the ten cluster centers (Figure 4(e)). In Figures 4(a), 4(b), 4(c) and 4(d)) the averages of the BoC corresponding to C_{10} of the classes computed tomography (CT), histopathology (HX), general photos (PX) and graphs (GX) are presented. We can observe that CTs (Figure 4(a)) are not only represented by black and white but also a few other colors. These colors are not represented stronger because several of the CT images in the database used are not fully grayscale images but RGB representations that have some color components. There are also color annotations on the images as they were used in journal texts. The HX BoC (Figure 4(b)) contains mainly red, green, pink and white colors, which is consistent with expectations. The PX BoC (Figure 4(c)) consists of a large variety of colors since it is a class with a very varied content. In the last example, we observe that the GX BoC 4(c) includes mostly white and some black, which is also consistent with expectations.

[4] http://fiji.sc/



(a) CT

(b) HX

(c) PX

(d) GX

(e) $C_{10} = \{c_1, \ldots, c_{10}\}$.

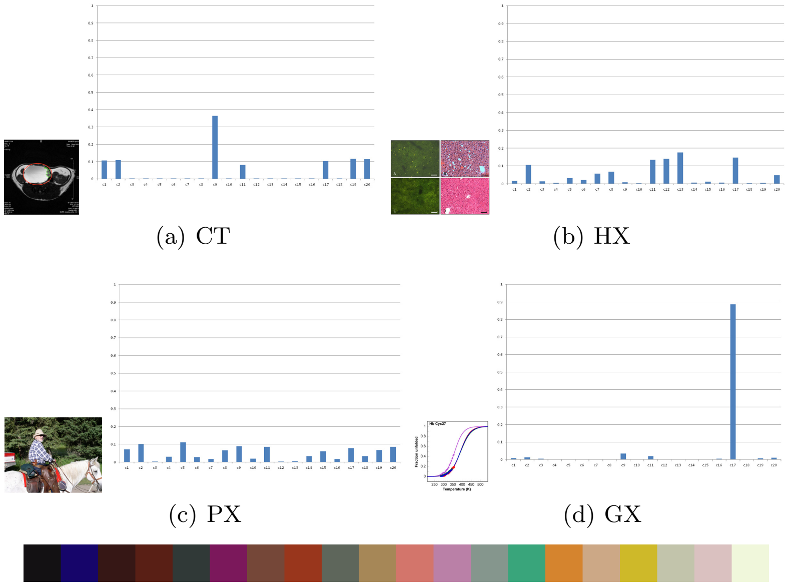
Fig. 4. Average BoC for four modalities corresponding to the color vocabulary $C_{10} = \{c_1, \ldots, c_{10}\}$ (4(e)) learned with $k_c = 10$

Given $k_c = 20$, a vocabulary $C_{20} = \{c_1, \ldots, c_{20}\}$ is shown in Figure 5(e) with the same examples. Since there are more colors each modality is represented by a BoC with a larger variety, follow similar patterns as with C_{10}.

The results for the training data with varying k_c over BoC and k_{nn} are shown in Table 1. And the results for the test data with varying k_{nn} are shown in Table 2. We use these results to choose the parameters. Since the selection of number of clusters affects the result of accuracy during image retrieval, five numbers of clusters were chosen: 10, 20 ,200, 500 and 1,000. Results indicate that classification performance has been improved by using 200 clusters.

We applied the optimal vocabulary size $k_c = 200$ on the test data As seen in the confusion matrices in Figure 6, there are more misclassified color images using SIFT than BoC such as in histopathology (HX), general photos (PX) or fluorescence (FL). On the other hand, using SIFT, there are fewer mistakes in radiology images (grey level) such as magnetic resonance imaging (MR), angiography (AN) or x–ray (XR). Figure 6(c) shows that the fusion of SIFT and BoC reduces the number of errors in both, color and typically grey level image types.

Using only SIFT, the best accuracy is 62.5% and results are stable for varying k_{nn}. For BoC the best accuracy is 63.96%, also quite stable across varying k_{nn}. For each k_{nn}, the fusion of BoC and SIFT produces an improved accuracy. The best overall fused result is 72.46%.



(a) CT

(b) HX

(c) PX

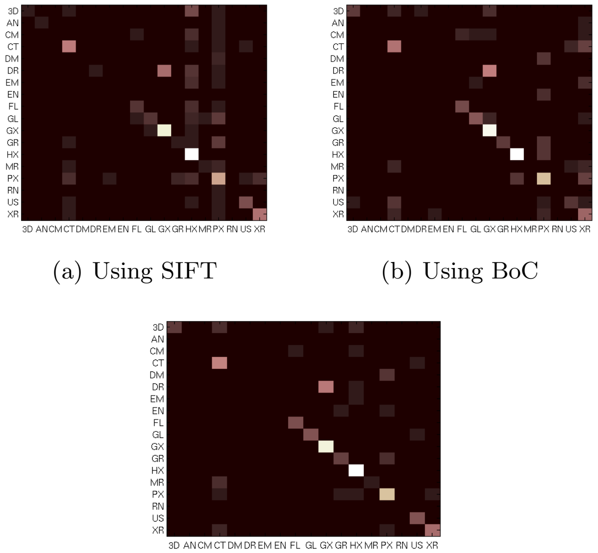
(d) GX

(e) $C_{20} = \{c_1, \ldots, c_{20}\}$.

Fig. 5. Average BoC for four modalities corresponding to the color vocabulary $C_{20} = \{c_1, \ldots, c_{20}\}$ (4(e)) learned with $k_c = 20$

Table 1. Classification accuracy using BoC/SIFT/both with varying k_c and k_{nn} over the training data

k_{nn}	SIFT	BoC & k_{10}	BoC & k_{20}	BoC & k_{200}	BoC & k_{500}	BoC & k_{1000}
2	27.27	14.83	24.40	24.83	24.84	24.40
3	28.06	17.80	27.58	26.81	26.81	27.03
4	28.56	16.15	28.13	28.46	28.46	28.24
5	28.85	17.80	27.80	28.90	29.23	28.46
6	29.15	17.14	28.57	**30.11**	29.45	30
7	28.95	19.01	28.57	29.78	29.34	28.57
8	28.85	19.01	29.78	**30.11**	29.01	28.68
9	**29.45**	18.57	29.67	28.68	28.79	28.79
10	29.35	18.35	29.45	29.12	29.23	28.35
11	29.05	18.57	28.79	29.12	29.12	29.23
12	29.15	18.79	28.90	29.01	29.89	29.01
13	28.75	18.79	29.34	29.12	28.68	29.01
14	28.95	18.57	29.23	29.45	28.79	28.68
15	29.35	18.79	29.45	28.79	28.46	28.35
16	29.25	18.02	28.90	28.35	28.68	28.35
17	**29.45**	18.35	28.79	27.91	28.57	28.79
18	29.25	18.57	29.45	27.80	28.46	27.91
19	29.15	18.35	29.56	27.91	28.46	27.80



(a) Using SIFT

(b) Using BoC

(c) Using SIFT and BoC

Fig. 6. Confusion Matrices obtained for the classification results using three feature types

Table 2. Classification accuracy using BoC/SIFT/both with varying k_{nn} on the test data

k_{nn}	SIFT	BoC	SIFT+BoC	k_{nn}	SIFT	BoC	SIFT+BoC
2	59.77	54.20	63.96	11	61.23	63.28	**72.46**
3	60.94	59.47	69.14	12	60.94	63.09	71.58
4	62.01	59.96	70.61	13	61.23	62.40	72.17
5	62.21	62.60	71.39	14	61.52	63.18	71.48
6	**62.50**	62.99	70.61	15	61.04	63.18	70.61
7	62.40	62.60	71.19	16	60.55	**63.96**	70.70
8	**62.50**	63.48	71.58	17	61.04	63.67	70.51
9	61.82	62.89	71.29	18	60.64	63.38	70.41
10	61.62	63.48	70.61	19	60.16	63.67	70.12

The k_{nn} voting is a very simple but often powerful tool [33]. We looked for the optimal k_{nn} value using the accuracy on the training data (Table 1). We also showed several k values on the training and test data to show the relative stability of the results. Furthermore, we tested Support Vector Machines (SVM) for our experiments. We used the Gaussian Radial Basis Function (RBF) kernel provided by WEKA[5] optimizing the parameters over the test set. The results were not as good as k_{nn} (Table 3), probably due to the characteristics of the database or the distribution of the features used.

[5] http://www.cs.waikato.ac.nz/ml/weka/

Table 3. Classification accuracy using BoC/SIFT/both using a simple SVM over the training data

Features	SIFT	BoC	SIFT+BoC
Accuracy	15.92	63.09	18.95

The best result in the modality classification task of ImageCLEF 2011 using visual methods [23] was obtained by Xerox research with 83.59% accuracy. This result is not comparable with our technique as the improvement was mainly due to an increased training set using data other than the original training data. Without the additional training set the obtained performance was at only 62.2% [34]. The best accuracy using visual methods without increasing the training data was 69.72%, obtained by the University of Campinas [35]. Using our fusion strategy of BoC and SIFT a better accuracy was obtained.

4 Conclusions and Future Work

In this paper, we present a BoC model for biomedical image categorisation. This domain is important for applications that aim at the integration of image retrieval tools into real applications We showed that fusing BoC and SIFT leads to good results for classifying biomedical document images. Results obtained by this approach demonstrate the notable improvement using BoC and SIFT together and also compared to 15 other research groups participating in ImageCLEF who work on the same data and in the exact same evaluation setting. There are other classification scheme like the one proposed by in [36]. We chose ImageCLEF because it has established standar database where we could compare aproches. However,the classes and ground truth provided by ImageCLEF are quite limited, ambiguous and, hence, reduces the quality of results obtaines.

Several directions are foreseen for future work. We plan to increase the training set for improved results as shown by Xerox in the competition in 2011. With a database available that is much larger than the training and test data sets this should be easily feasible. Using a different color space or adding additional features can be another option but can be expected to lead to only small improvements. Particularly the text captions can be used for classification improvement as some classes can easily be distinguished using the captions. Such mixtures of visual and textual methods equally have a potential for important performance improvements.

Acknowledgments. This work was partially funded by the EU in the context of the KHRESMOI (257528) and PROMISE (258191) FP7 projects.

References

1. Hunter, L., Cohen, K.B.: Biomedical language processing: What's beyond pubmed? Molecular Cell 21(5), 589–594 (2006)
2. Depeursinge, A., Duc, S., Eggel, I., Müller, H.: Mobile medical visual information retrieval. IEEE Transactions on Information Technology in BioMedicine 16(1), 53–61 (2012)

3. Hersh, W., Jensen, J., Müller, H., Gorman, P., Ruch, P.: A qualitative task analysis for developing an image retrieval test collection. In: ImageCLEF/MUSCLE Workshop on Image Retrieval Evaluation, Vienna, Austria, pp. 11–16 (2005)
4. Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals' image use and search behaviour. In: Proceedings of the Medical Informatics Europe Conference (MIE 2006). Studies in Health Technology and Informatics, pp. 24–32. IOS Press, Maastricht (2006)
5. Hersh, W.R., Hickam, D.H.: How well do physicians use electronic information retrieval systems? Journal of the American Medical Association 280(15), 1347–1352 (1998)
6. Hoogendam, A., Stalenhoef, A.F., de Vries Robbé, P.F., Overbeke, A.J.: Answers to questions posed during daily patient care are more likely to be answered by uptodate than pubmed. Journal of Medical Internet Research 10(4) (2008)
7. Kahn, C.E., Thao, C.: Goldminer: A radiology image search engine. American Journal of Roentgenology 188(6), 1475–1478 (2007)
8. Rafkind, B., Lee, M., Chang, S.-F., Yu, H.: Exploring text and image features to classify images in bioscience literature. In: Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis, New York, NY, USA, pp. 73–80 (2006)
9. Demner-Fushman, D., Antani, S., Siadat, M.R., Soltanian-Zadeh, H., Fotouhi, F., Elisevich, K.: Automatically finding images for clinical decision support. In: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, ICDMW 2007, pp. 139–144. IEEE Computer Society, Washington, DC (2007)
10. Pentland, A.P., Picard, R.W., Scarloff, S.: Photobook: Tools for content–based manipulation of image databases. International Journal of Computer Vision 18(3), 233–254 (1996)
11. Lakdashti, A., Moin, M.S.: A New Content-Based Image Retrieval Approach Based on Pattern Orientation Histogram. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 587–595. Springer, Heidelberg (2007)
12. Jain, A.K., Vailaya, A.: Image retrieval using color and shape. Pattern Recognition 29(8), 1233–1244 (1996)
13. van de Sande, K.E., Gevers, T., Snoek, C.G.: A comparison of color features for visual concept classification. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, CIVR 2008, pp. 141–150. ACM, New York (2008)
14. Tou, J.Y., Tay, Y.H., Lau, P.Y.: Recent trends in texture classification: A review. In: Symposium on Progress in Information & Communication Technology, Kuala Lumpur, Malaysia, pp. 63–68 (2009)
15. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognition 37(1), 1–19 (2004)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
17. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. Compututer Vision and Image Understanding 113(1), 48–62 (2009)
18. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1582–1596 (2010)
19. Ai, D., Han, X.H., Ruan, X., Chen, Y.W.: Adaptive color independent components based sift descriptors for image classification. In: ICPR, pp. 2436–2439. IEEE (2010)

20. Markonis, D., García Seco de Herrera, A., Eggel, I., Müller, H.: Multi–scale visual words for hierarchical medical image categorization. In: SPIE Medical Imaging 2012: Advanced PACS–based Imaging Informatics and Therapeutic Applications, vol. 8319, pp. 83190F–11 (February 2012)
21. Wengert, C., Douze, M., Jégou, H.: Bag–of–colors for improved image search. In: Proceedings of the 19th ACM International Conference on Multimedia, MM 2011, pp. 1437–1440. ACM, New York (2011)
22. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence 27(10), 1615–1630 (2005)
23. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum) (September 2011)
24. Sharma, G., Trussell, H.J.: Digital color imaging. IEEE Transactions on Image Processing 6(7), 901–932 (1997)
25. Banu, M., Nallaperumal, K.: Analysis of color feature extraction techniques for pathology image retrieval system. IEEE (2010)
26. Grauman, K., Leibe, B.: Visual Object Recognition (2011)
27. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
28. Hinneburg, A., Keim, D.A.: An efficient approach to clustering in large multimedia databases with noise. In: Conference on Knowledge Discovery and Data Mining (KDD), vol. 5865, pp. 58–65. AAAI Press (1998)
29. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
30. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: MULTIMEDIA 2005: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402. ACM, New York (2005)
31. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: Text REtrieval Conference, pp. 243–252 (1993)
32. Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining (Adaptive Computation and Machine Learning). The MIT Press (2001)
33. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
34. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In: Working Notes of CLEF 2011 (2011)
35. Faria, F.A., Calumby, R.T., da Silva Torres, R.: RECOD at ImageCLEF 2011: Medical modality classification using genetic programming. In: Working Notes of CLEF 2011 (2011)
36. Deserno, T.M., Antani, S., Long, L.R.: Content–based image retrieval for scientific literature access. Methods of Information In Medicine 48(4), 371–380 (2009)

An SVD–Bypass Latent Semantic Analysis for Image Retrieval

Spyridon Stathopoulos and Theodore Kalamboukis

Department of Informatics, Athens University of Economics and Business,
Athens, Greece
spstath@gmail.com, tzk@aueb.gr

Abstract. This article presents an experimental evaluation on image representation using Latent Semantic Analysis (LSA) for searching very large image databases. Our aim is twofold: First, we experimentally investigate the structure and size of the feature space in order for LSA to bring efficient results. Second, we replace the Singular Value Decomposition (SVD) analysis on the feature matrix, by solving the eigenproblem of the term correlation matrix, a much less memory demanding task which significantly improved the performance in both accuracy and computational time (preprocessing and query response time) on three large image collections. Finally the new approach overcomes the high cost of updating the database after new insertions.

Keywords: LSA, LSI, CBIR.

1 Introduction

From experimental research, [1], it is evident that performance in image retrieval still is very far from being effective for several reasons: computational cost, scalability and performance. In this respect, several approaches have been investigated that combine visual retrieval with textual retrieval using annotations or text associated to images. An approach to reduce computational cost is achieved by dimensionality reduction, a process which has been extensively used in text retrieval, well known by the term Latent Semantic Indexing [2] (LSI). The aim of this work is to investigate the application of the LSA approach to image retrieval. LSA is a technique used to remove redundancy information and noise from data. It is a mathematical technique for extracting relations of contextual usages of words in documents. The use of LSA is herein proposed to discover latent associations between visual patterns in the feature space. In order for the LSA approach to give good results the following conjectures should apply [3]:

1. The features should be invariant to affine transformations.
2. The discriminating power of the feature set should be high. Features with high frequency or features that occur rarely within an image collection should be removed.
3. The features should mimic the human perception of similarity.

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 122–132, 2013.

4. The feature set size should be very high. In text retrieval the documents are represented as points in a potentially very high dimensional metric vector space.

The first step of SVD is to build the $features \times images$ $(m \times n)$ matrix C. An entry, $C[i, j]$, contains the weight of feature i in the image j. Applying SVD, the matrix C, is analysed into a product of three matrices $C = U\Sigma V^T$, where U $(m \times m)$ is the matrix of the right eigenvectors of CC^T, V^T, $(n \times n)$ is the matrix of the left eigenvectors of $C^T C$ and Σ is the diagonal matrix with $r \neq 0$ singular values ($r = rank(C)$) [4]. Using the eigenvectors corresponding to the k largest eigenvalues we build a new subspace of dimension k to represent the images.

LSA has been used in CBIR primarily on small data sets. The main reason for this is the complexity of SVD on large feature matrices. From our experiments, we conclude that a moderate number of features ($m \approx 10000$) is sufficient for effective retrieval with LSA. Based on this observation, we propose a "bypass" solution to the SVD problem which overcomes all its deficiencies, concerning time and space requirements and makes the method attractive for content-based image retrieval. In the next section, we describe our approach and in the following sections we proceed with the implementation details and extensive experiments on three image collections. Finally the last section contains the concluding remarks of this work with propositions for future research.

2 Related Work

Although LSA has been used successfully in many applications in the domain of information retrieval, it has not experienced similar success in CBIR. This is due to the fact that in the case of IR, the matrix C is sparse and therefore it can be stored in compact form (only non-zero elements are stored). In this case, even if there is not enough space to accommodate the matrix in memory there are effective iterative solvers such as Krylov subspace methods [4]. In contrast, in CBIR, the features' matrix C, is dense and this raises the complexity cost of SVD to prohibitively high levels for both, space and computational time. We refer here to the work in [5,6,7] that use LSA in CBIR and represent images by vectors of size $x \cdot y$ where $x \cdot y$ represents the number of pixels in the entire image. An entry $C[i, j]$ represents the color in the i-th position in the j-th image. In the example given in [5] the database contains 71 images of size 640×480 pixels which produces a matrix C, of order 307200×71. The authors instead of solving the SVD for the matrix C, solve the eigenproblem for $C^T C$, a matrix of order 71×71. However, the size of the collection is very small and the problem does not scale to large image collections.

In this work we propose a similar and still effective and scalable application of LSA in Content-Based Image Retrieval. Instead of solving the SVD for the matrix C, we solve the eigenproblem of the $m \times m$ correlation matrix CC^T where m has a moderate value between $5000 - 10000$. This problem demands only $O(m^2)$ space and can be solved easily using several software packages available on the web such as, Scilab, JAMA library or Matlab. A weakness of this approach is that it may lead to instability of eigenvectors in terms of the condition number of the matrix CC^T. However, as is known the separations of eigenvalues have an influence on the sensitivity of eigenvectors [4]. If the eigenvalues are well separated then the corresponding eigenvectors are orthogonal, provided that they give a small residual norm and that was the case in all our experiments. Furthermore our experimental results confirm that our approach combined with with image splitting improves significantly the performance over the MPEG-7 color descriptors we have tested.

3 Description of the Algorithm

According to the traditional use of LSA in information retrieval a *term-by-document* matrix, C, is first constructed and SVD analysis is performed on this matrix. Next, a low rank approximation of the matrix C is applied by $C_k = U_k \Sigma_k V_k^T$, for $(k << r)$ keeping the eigenvectors corresponding to the k largest eigenvalues. The documents in the new k-dimensional space are represented by the columns of the matrix V^T and the queries are projected in this space by: $\hat{q} = \Sigma_k^{-1} U_k q$ [2]. The similarity of a query with a document is calculated by the cosine function in the reduced space. This is equivalent to solving the eigenproblem for the matrix CC^T as it is shown in the following lemma.

Lemma
Given a matrix C $(m \times n)$ of rank r, and the right eigenvectors, U, with the corresponding eigenvalues Σ^2 of the term-correlation matrix CC^T it holds that $V^T = \Sigma^{-1} U^T C$ is an orthonormal matrix that forms the left eigenvectors of $C^T C$ and $C = U \Sigma V^T$.

Proof
It is straightforward to show that the set of vectors $V^T = \Sigma^{-1} U^T C$ are orthonormal $(V^T V = I)$ and they form the left eigenvectors of the matrix $C^T C$, $(V^T C^T C = \Sigma^2 V^T)$. Finally, since the matrices U and V are orthonormal it holds that $C = U \Sigma V^T$. Indeed $U \Sigma V^T = U \Sigma (\Sigma^{-1} U^T C) = C$
Our approach: A "bypass" SVD Analysis

Our LSA implementation is summarized as follows:

1. Solve the $m \times m$ eigenproblem $CC^T U = U \Sigma^2$, select the k largest eigenvectors and corresponding eigenvalues
2. Calculate the projections of the original images into the k-th dimensional space $\hat{d}_j = U_k^T d_j$
3. For a query q, calculate the similarity $score(\hat{q}, \hat{d}_j)$, by the cosine function, where $\hat{q} = U_k^T q$

In the case of image retrieval, when we decide on the selection and spatial information of the low level features, images are represented by feature vectors which are dense, very large but beforehand all have a fixed size. For example, if an image is split into 324 tiles, as will see in the following sections for one of our databases, using a grayscale histogram with 32 bins each image is represented by a dense vector of size $324 \times 32 = 10368$. Thus for the whole data collection consisting of 231000 images the feature matrix needs about 19GB of memory. Things get worse when images are represented by full pixel data. For example if images consist of 216×216 pixels our database will require 86GB of memory. For such image collections it was impossible to solve the SVD problem on our computer, a Pentium-i5 with 4GB RAM. With our "SVD-bypass" approach we have to solve an eigenproblem for a matrix of order 10368×10368 for the k=50 largest eigenvalues and corresponding eigenvectors which is not only possible but a very fast task on our machine. Another important advantage of our new approach compared to the traditional LSA procedure is its efficiency with updates of the database. Indeed if a new set of images represented, say, by a matrix C_n, is added, we calculate the matrix $C_nC_n^T$ and solve an eigenproblem of the same order $(m \times m)$ for the matrix $CC^T := CC^T + C_nC_n^T$. Concerning the stability of the eigensolution for the matrix CC^T we have observed in all our experiments that the largest eigenvalues are well separated from each other and the best value of k was near the elbow of the corresponding scree plot (see Figure 1).

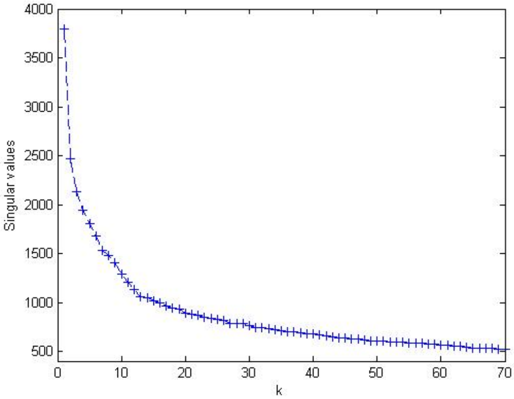


Fig. 1. The distribution of the largest singular values of the feature matrix for Corel-5k (k=20)

4 Preprocessing of the Data

It is well known that the representation of a digital image depends on several factors, from its resolution to color models etc. In a collection of images, it is highly possible that there will be important variations considering these characteristics. Thus, each image undergoes through several transformations before the feature extraction step. In our image collections we have applied the following transformations.

Size Normalization of Images. The purpose of rescaling the images in a database is to normalize the feature vectors. In our image collections a rescaling consisting of 216 × 216 pixels in the RGB color space was chosen to compute the feature vectors. This choice is crucial especially in our medical image collections were images from different modalities have a different resolution. Bilinear interpolation[1] was used for the rescaling process. The method overlays a grid of 216 × 216 points on the input image and the colors of the pixels at these grid point in the output image are determined by linearly interpolating between adjacent pixel colors (both vertically and horizontally).

Transform to Grayscale. The images were converted into a 256 value grayscale model. This transformation was chosen primarily in order to reduce the computational complexity as well as for color normalization since our collections contained both color and gray scale images. Although converting a digital image from color to grayscale is considered to be a simple process, certainly there are more sophisticated ways that may deliver better results. In our case we used for this task an open source fast and efficient Java method[2] found in the internet.

Splitting Images into Tiles. Our main concern is to reduce the size of the feature vectors by splitting the images into tiles and extracting low level features from each tile. Image tiling means the splitting of a given image into non-overlapping cells or squares of equal size. Here we have considered 16, 32, 64, 324 non overlapping tiles as shown in Figure 2. For example, in the case of splitting an image into 324 tiles each tile is a square of 18 × 18 pixels. The final vector representation of an image is obtained by concatenating the feature vectors of the tiles row-wise. The size of the feature vectors for images of 16 tiles for the Color Layout MPEG-7 descriptor will be equal to $16 * 192 = 3072$. This process of tiling at the limit will represent images at pixel level and the size of the feature vectors will become m^2. From our experiments we observed that the number of tiles cannot grow indefinitely and beyond a number they have a negative effect.

Feature Extraction and Selection. The low level features (scalable color and color layout) of MPEG-7 were extracted using the library Caliph&Emir of the Lire CBIR system [8]. With the color layout descriptor each image is represented by a vector of 192 coefficients and for the scalable color descriptor with 64 coefficients. For the color histogram, a method was implemented which extracts three histograms in the case of an RGB image or one histogram in the case of a grayscale image. As it is implied from conjecture 2, in the introduction, the extracted features should have a high discriminating power. Therefore for each feature j the mean and standard deviation was calculated for the whole image set and new weights were estimated for each feature, j, in the image vector by:

$$w_j = \begin{cases} \mid \frac{c_{ij}-\mu_j}{\sigma_j} \mid & \text{if } c_{ij} > \mu_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

[1] http://en.wikipedia.org/wiki/Bilinear_interpolation

[2] http://www.codebeach.com/2008/03/convert-color-image-to-gray-scale-image.html

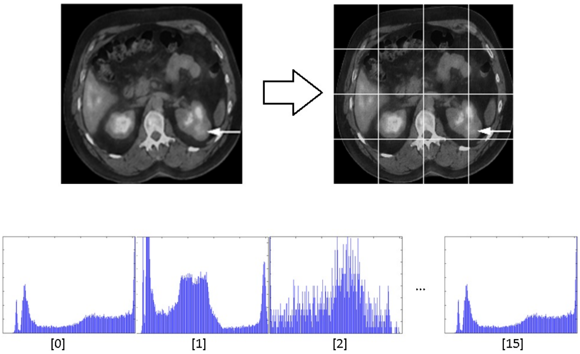


Fig. 2. Segmentation of an image into 16 tiles. The image is represented by the color histograms of each tile.

The goal is to remove the features with high frequency and normalize the remaining feature values for all images. At the same time, the frequency for each feature j was calculated and the features with a frequency above 80% are considered stopwords and thus, they were removed.

Construction of the Feature-Correlation Matrix $\boldsymbol{CC^T}$. As we have already mentioned the matrix C is full in the case of CBIR and so it is the matrix CC^T. This matrix multiplication is the most intensive part of the computations and memory demanding. In our implementation we overcome all these problems by splitting the matrix C into a number of blocks, such that each block can be accommodated in memory ($C = (C_1, C_2, ..., C_p)$) and calculate CC^T by:

$$CC^T = \sum_{i=1}^{p} C_i C_i^T \tag{2}$$

5 Databases and Experimental Results

5.1 Databases

Accessing a large image collection and obtaining useful information is a difficult task and an active research area with very poor effectiveness compared to text retrieval [1]. Throughout our tests we have used the image collections from the imageCLEF[3] Ad-Hoc task over the past years (2010-2011). Such image collections may be used to support the decision making process in clinical services, as well as for research purposes and education. These medical collections, however, are very generic and cover a large spectrum of images, from MRI, to x-rays and diagrams. Since the performance of content-based retrieval in such collections without the use of context (image captions) is very poor, we have included in our experiments the Corel-5k collection, a common set for benchmarking in CBIR.

[3] http://www.imageclef.org/

ImageCLEF-2010 database contains a total of 77,506 images [1]. The image-based topics were created using realistic methods by the Radiological Society of North America and they were identified by surveying actual user needs. Sixteen queries were selected for ImageCLEFmed-2010. Each query contains a visual topic which is associated with 2 to 4 sample images from other collections of ImageCLEFmed and a textual topic with a very short description.

ImageCLEF-2011 collection is a subset of 231,000 images from the PubMed Central database [9]. Similarly, in 2011, thirty topics were selected from these which retrieved at least one relevant image by the system.

Corel-5k dataset has been widely used in the literature to evaluate image retrieval and classification systems [10]. It consists of $5,000$ images from 50 thematic categories such as images of tigers, images of cars etc. The collection is divided into a training set of 4,500 images and a test set of 500 images where in the test set there are 10 images from each thematic category. Retrieval performance is evaluated using the thematic category information of each query.

We have run three sets of experiments. The first set refers to retrieval using color descriptors on the original images. Due to restrictions on computing resources we have run a second set of experiments on the same collections with images converted to grayscaled and rescaled to 216×216 pixels each. These two sets of experiments form the ground truth results (Tables 1, 2) for comparison with our LSA approach in the third set of experiments.

Table 1. Ground truth results of color descriptors on the Corel-5k image collection

	Original Images				Grayscale - Resize of images			
	Scalable	Color	Color	Color	Scalable	Color	Color	Color
	Color	Layout	Histogram	Selection	Color	Layout	Histogram	Selection
MAP	0.0510	0.0429	0.0283	0.0173	0.0365	0.0136	0.0235	0.0070
P@5	0.1439	0.1399	0.1042	0.0665	0.0373	0.0333	0.0970	0.0200
P@10	0.1405	0.1327	0.0996	0.0669	0.0445	0.0403	0.0962	0.0200
P@20	0.1258	0.1193	0.0912	0.0607	0.0482	0.0418	0.0867	0.0200
Rel_ret	19263	17379	15466	13061	19584	12801	14029	8950

In the case of a multi-image visual topics (more than one image per topic) the simple CombSUM scoring function is used,defined by:

$$SCORE(Q, Image) = \sum_{im_i \in Q} score(im_i, Image) \tag{3}$$

where $Q = \{im_1, ..., im_p\}$ is a set of p query-imagies im_i. For the evaluation we follow the TREC example using the software tool TREC_EVAL[4] on the first 1000 retrieved images. The evaluation is based on the mean average precision (MAP)[5] value and the precision at the top k retrieved items. For the CLEF

[4] http://trec.nist.gov/trec_eval/

[5] http://en.wikipedia.org/wiki/Information_retrieval

Table 2. Ground truth results of color descriptors on the clef 2010 and 2011 image collections

		Original Images				Grayscale - Resize of images			
		Scalable Color	Color Layout	Color Histogram	Color Selection	Scalable Color	Color Layout	Color Histogram	Color Selection
clef 2010	MAP	0.0006	0.0006	0.0033	0.0005	0.0002	0.0004	0.0005	0.0001
	bpref	0.0241	0.0174	0.0104	0.0071	0.0109	0.0136	0.0084	0.0063
	P@5	0.0000	0.0000	0.0125	0.0125	0.0000	0.0000	0.0000	0.0000
	P@10	0.0000	0.0000	0.0063	0.0063	0.0000	0.0000	0.0063	0.0000
	P@20	0.0031	0.0000	0.0031	0.0031	0.0000	0.0000	0.0031	0.0031
	Rel_ret	30	32	25	13	20	34	21	13
clef 2011	MAP	0.0009	0.0062	0.0009	0.0006	0.0002	0.0013	0.0016	0.0000
	bpref	0.0206	0.0529	0.0322	0.0304	0.0192	0.0373	0.0377	0.0005
	P@5	0.0067	0.0533	0.0000	0.0000	0.0000	0.0067	0.0067	0.0000
	P@10	0.0067	0.0500	0.0067	0.0000	0.0000	0.0100	0.0133	0.0000
	P@20	0.0083	0.0333	0.0050	0.0033	0.0000	0.0083	0.0083	0.0000
	Rel_ret	120	423	162	131	54	208	212	3

databases we calculate also the binary preference (bpref) [11] a more robust metric in the case of incomplete judgements.

From Tables 3 and 4 we see that LSA performs significantly better for all color descriptors compared to the ground truth results in Table 1. Best performance in terms of MAP was attained for scalable color with 64 tiles and low rank approximation of $k = 20$ was enough to capture the feature relationship information in the collection. This experiment corresponds to the solution of an eigenproblem of size 4096×4096. The largest eigenproblem for the Corel dataset was 12288×12288 corresponded to the color layout descriptor with 64 tiles.

Table 3. LSA with Color Histogram on Corel-5k collection with 16 and 32 grayscale colors and k=20, 50

Corel-5k	Tiles=36		Tiles=324	
colors=16 k=20	Color Histogram	Color Selection	Color Histogram	Color Selection
MAP	0.0220	0.0209	0.0258	0.0211
P@5	0.0922	0.0798	0.1102	0.0790
P@10	0.0912	0.0762	0.1046	0.0804
P@20	0.0791	0.0691	0.0902	0.0722
Rel_ret	14448	14326	15339	14501
colors=32 k=50	Tiles=36		Tiles=324	
MAP	0.0244	0.0214	0.0295	0.0230
P@5	0.1062	0.0858	0.1175	0.0942
P@10	0.1012	0.0812	0.1132	0.0930
P@20	0.0835	0.0726	0.0955	0.0792
Rel_ret	14678	14264	15927	14519

Table 4. LSA of color descriptors (scalable color and color layout) on core-5k collection

	Corel-5k, Tiles=36				Corel-5k, Tiles=64			
k	20		50		20		50	
	Scalable Color	Color Layout	Scalable Color	Color Layout	Scalable Color	Color Layout	Scalable Color	Color Layout
MAP	0.0573	0.0516	0.0554	0.0519	0.0589	0.0521	0.0569	0.0524
P@5	0.1655	0.1723	0.1659	0.1739	0.1711	0.1731	0.1627	0.1707
P@10	0.1595	0.1657	0.1579	0.1613	0.1593	0.1665	0.1621	0.1627
P@20	0.1455	0.1480	0.1422	0.1454	0.1441	0.1484	0.1441	0.1444
Rel_ret	18407	18772	17382	18579	18870	18816	17641	18686

Table 5. LSA of color descriptors on clef-2010 and 2011 collections

k=50		Tiles=36				Tiles=64			
		Scalable Color	Color Layout	Color Histogram	Color selection	Scalable Color	Color Layout	Color Histogram	Color selection
clef 2010	MAP	0.0005	0.0034	0.0005	0.0020	0.0007	0.0036	0.0005	0.0028
	P@5	0.0000	0.0125	0.0000	0.0000	0.0000	0.0250	0.0000	0.0125
	P@10	0.0000	0.0063	0.0000	0.0000	0.0063	0.0188	0.0000	0.0063
	P@20	0.0000	0.0063	0.0000	0.0094	0.0031	0.0094	0.0000	0.0125
	Rel_ret	12	47	46	46	16	42	45	47
clef 2011	MAP	0.0034	0.0134	0.0026	0.0016	0.0042	0.0135	0.0027	0.0017
	P@5	0.0400	0.1067	0.0200	0.0067	0.0400	0.0933	0.0267	0.0067
	P@10	0.0233	0.0800	0.0333	0.0100	0.0267	0.0833	0.0233	0.0100
	P@20	0.0167	0.0700	0.0250	0.0100	0.0183	0.0750	0.0217	0.0117
	Rel_ret	151	537	292	219	156	542	283	233

The same behaviour with LSA was shown on both CLEF databases (Table 5). Here the best value of k was taken equal to 50.

5.2 Data Fusion

For the data fusion task we used a weighted linear combination of the results obtained from the color descriptors, as defined by :

$$SCORE(Q, Image) = \sum_{i} w_i * score_i(Q, Image) \tag{4}$$

were $score_i$ denotes the similarity score of an $Image$ with respect to a color descriptor i and the weights w_i are expressed as functions of the value of MAP attained by the corresponding descriptor. Table 6 presents the results of fusion of the descriptors, Dominant Color Histogram, Scalable Color and Color Layout and LSA with 64 tiles, and $k = 20$. For the CLEF databases $k = 50$ was used. The weights w_i were taken equal to MAP$_i$ squared.

Results in Table 6 are impressive for CBIR using only color descriptor. We note here that for a fair comparison, the results of LSA should be compared to the second column in Table 1 entitled ”grayscale”.

Table 6. Data fusion with Dominant Color Histogram, Scalable Color and Color Layout

	Corel-5k			clef-2010			clef-2011		
	Original	Grayscale	LSA	Original	Grayscale	LSA	Original	Grayscale	LSA
MAP	0.0618	0.0353	**0.0519**	0.0008	0.0003	**0.0017**	0.0060	0.0013	**0.0106**
P@5	0.2000	0.0741	**0.1727**	0.0000	0.0000	**0.0125**	0.0533	0.0067	**0.0733**
P@10	0.2012	0.0806	**0.1567**	0.0000	0.0000	**0.0063**	0.0467	0.0100	**0.0533**
P@20	0.1709	0.0758	**0.1349**	0.0000	0.0000	**0.0031**	0.0317	0.0083	**0.0444**
Rel_ret	19794	19582	**18883**	27	28	**41**	420	208	**542**

6 Conclusions

We have presented a new approach to LSA for CBIR replacing the SVD analysis of the feature the matrix C $(m \times n)$ by the solution of the eigenproblem for the matrix CC^T $(m \times m)$. From all our experiments in three image collections follows that m is of a moderate size between 5000 to 10000 and $m << n$. From all these experiments it is apparent that the proposed method is much faster, requires less memory, is robust and outperforms standard methods on the same low level features consistently in all our experiments. The method overcomes the high cost of SVD to update the database when new images are inserted. In addition, in all the experiments the optimal value of the approximation parameter was less than 50 which makes the method attractive for fusion with several low level features. Our results from data fusion on color descriptors show that LSA outperforms significantly similar methods on the same descriptors [8].

Certainly our approach is promising and has created new research directions that need further investigation. The image representation has an impact on LSA performance and a more systematic research on that direction is currently under progress. Another challenge that needs further investigation concerns the stability of the eigenproblem for the matrix CC^T and the behavior of the problem when n grows to infinity. So far in all our examples, the best performance of the LSA was achieved for very small values of $k \approx 50$. Also in all the cases the k-th largest singular values were well separated giving small residual vectors to machine accuracy, which give us the evidence on the stability of the calculated eigenvectors.

References

1. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Reisetter, J., Kahn, C.E., Hersh, W.R.: Overview of the clef 2010 medical image retrieval track. In: CLEF (Notebook Papers/LABs/Workshops) (2010)

2. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS 41(6), 391–407 (1990)
3. Pecenovic, Z., Lausanne, H., Losanna, P.F., Dra, F., Lau, L.D.E., Anne, S., Dr, A., Ayer, S., Vetterli, P.M.: Image retrieval using latent semantic indexing (1997)
4. Golub, G.H., van Loan, C.F.: Matrix computations, 3rd edn. Johns Hopkins University Press (1996)
5. Skopal, T., Snásel, V.: An application of lsi and m-tree in image retrieval. GESTS International Transaction on Computer Science and Engineering 34(1), 212–223 (2006)
6. Praks, P., Snasel, V., Dvorsky, J., Cernohorsky, J.: On svd-free latent semantic indexing for image retrieval for application in a hard industrial environment. In: IEEE International Conference on Industrial Technology, vol. 1, pp. 466–471. Conference Publications (2003)
7. Praks, P., Kucera, R., Izquierdo, E.: The sparse image representation for automated image retrieval. In: Proceedings of the International Conference on Image Processing, ICIP 2008, San Diego, California, USA, October 12-15, pp. 25–28 (2008)
8. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: El-Saddik, A., Vuong, S., Griwodz, C., Bimbo, A.D., Candan, K.S., Jaimes, A. (eds.) ACM Multimedia, pp. 1085–1088. ACM (2008)
9. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsikrika, T.: Overview of the clef 2011 medical image classification and retrieval tasks. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
10. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
11. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Sanderson, M., Järvelin, K., Allan, J., Bruza, P. (eds.) SIGIR, pp. 25–32. ACM (2004)

Multimedia Retrieval in a Medical Image Collection: Results Using Modality Classes

Angel Castellanos[1], Xaro Benavent[2],
Ana García-Serrano[1], and J. Cigarrán[1]

[1] Universidad Nacional de Educación a Distancia, UNED
[2] Universitat de València
xaro.benavent@uv.es,
{acastellanos,agarcia,juanci}@lsi.uned.es

Abstract. The effective communication between user and systems is one main aim in the Multimedia Information Retrieval field. In this paper the modality classification of images is used to expand the user queries within the ImageCLEF Medical Retrieval collection provided by organizers. Our main contribution is to show how and when results can be improved by understanding modality-related challenges. To do so, a detailed analysis of the results of the experiments carried out is presented and the comparison between these results shows that the improvement using modality class query expansion is query-dependent.

Keywords: Information Retrieval, Text-based Retrieval, Content-Based Image Retrieval, Merge Results Lists, Fusion, Indexing.

1 Introduction

One of the main aims within current research in multimedia retrieval is the effective communication between users and the system. One further step is the use of the image modality class information in the multimedia search process. Previous studies have shown that the visual-modality information is relevant for Content-Based Information Retrieval (CBIR) in medical domain repositories [12]. Modality classification of images in a medical collection was used in different research in previous work [5, 13].

In recent years of ImageCLEF, different methods of query expansion have been proposed. The bulk of them focused on using external medical sources such as MeSH (Medical Subject Headings) [2] or UMLS (Unified Language Medical System) [14]. The expansion of queries with other external sources such as Wikipedia [6, 7] was raised.

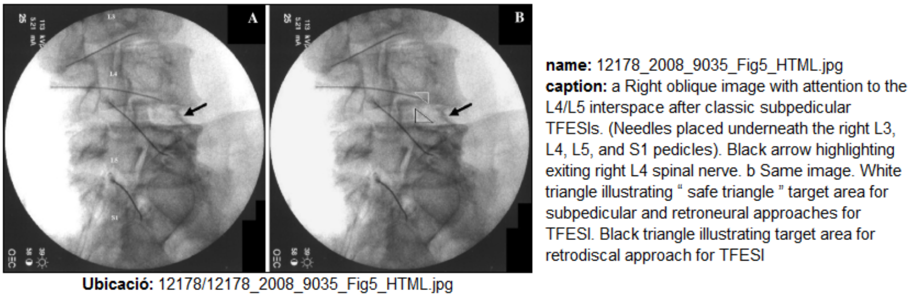
The main goal of the presented work is to investigate how to deal with the visual-modality information for query expansion in Text-Based Information Retrieval (TBIR) for Multimedia retrieval systems. Our approach is based on query expansion using the modality classification information for TBIR. The contribution of the present work is the analysis of how the results using the modality classification information vary depending on the query.

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 133–144, 2013.

In this paper, at section 2 it is presented the textual preprocessing information of the ImageCLEF Medical collection performed. A detailed overview of our own developed Multimedia Search System is presented at section 3. Then the experiments carried out, that range from mono-modal to multimodal ones, and a new set of experimentation and obtained results is included in section 4. This experimentation is complementary to the one presented by the authors at ImageCLEF [4]. Finally, in section 5 we extract conclusions and point out further work.

2 Textual Preprocessing of the ImageCLEF Medical Collection (2011)

In the ImageCLEF Medical Retrieval Task at the 2011 International Campaign [8], the image collection provided to the participants contains a dataset, queries and relevance judgments of the queries. This collection is made up of a dataset of 230,088 images from PubMed Central and is structured into medical articles. The images are described by textual descriptions (see Fig. 1).



name: 12178_2008_9035_Fig5_HTML.jpg
caption: a Right oblique image with attention to the L4/L5 interspace after classic subpedicular TFESIs. (Needles placed underneath the right L3, L4, L5, and S1 pedicles). Black arrow highlighting exiting right L4 spinal nerve. b Same image. White triangle illustrating " safe triangle " target area for subpedicular and retroneural approaches for TFESI. Black triangle illustrating target area for retrodiscal approach for TFESI

Ubicació: 12178/12178_2008_9035_Fig5_HTML.jpg

Fig. 1. System Interface with an image of the collection and associated description

The ImageCLEF Medical Retrieval Task organization also provided a modality classification of the images, along with the collection. The 18 categories are:

1. 3D: 3d reconstruction
2. AN: angiography
3. CM: compound figure
4. CT: computed tomography
5. DM: dermatology
6. DR: drawing
7. EM: electron-microscopy
8. EN: endoscopic imaging
9. FL: fluorescence
10. GL: gel
11. GX: graphs
12. GR: gross pathology
13. HX: histopathology
14. MR: magnetic resonance
15. PX: general photo
16. RN: retinograph
17. US: ultrasound
18. XR: x-ray

The whole collection was divided by the authors into several files. One xml file for each of the articles that makes up the collection is generated and then an additional field (called *tags*) is included in an automatically way. This field stores the modality of the image established by the classification provided by the organizers. The Text-based process can use or not this modality information to test the possible benefits of including this information to the query.

3 System Description

The overall system includes three main subsystems: the TBIR, the CBIR, and the Fusion subsystem (see Figure 2). Both the textual (TBIR) and the visual subsystem (CBIR) obtain a ranked list of images based on similarity scores (St and Si) for a given query. First, TBIR uses the textual information from the annotations to obtain these scores (St). This textual pre-filtered list is then used by the CBIR sub-system. This idea, which was used by the authors in previous edition of ImageCLEF [3], is based on the assumption that the conceptual meaning of a topic is initially better captured by the text module itself than by the visual module. Then the CBIR system extracts the visual information from the given example images of the query and generates a similarity score (Si). The fusion sub-system is in charge of merging these two lists of results, taking into account the scores and rankings, in order to obtain the final result list.

The **TBIR subsystem** prior to working with the collection carries out a preprocessing stage and the indexing of the images for the subsequent search. The result of the search from a query is an image list, which is ranked according to its similarity with the corresponding query in accordance only with the textual information.

The concrete stages carried out by the TBIR subsystem are:

1. **Collection Reformatted & Expansion** as explained in section 2.
2. **Query File Construction:** Based on the original query file, query files are constructed to address the different experiments (runs explained in the section 4).
3. **Preprocessing:** Textual information is preprocessed in different ways: 1) characters with no meaning, such as punctuation marks or blanks, are eliminated; 2) stopword deletion (words without semantic information) 3) stemming or reduction of words to their base form and finally 4) conversion of words to lower case.
4. **Indexing:** The indexing is carried out automatically by Solr[1] making use of the Lucene[2] algorithms. The time of indexing, about 7 minutes, is relatively short considering the size of the collection.
5. **Searching:** This process is started manually, running each query in the Solr interface. The results, returned in xml format, are transformed to the TRECEval format, in order to merge these textual results with visual results, obtained by the CBIR subsystem, and thus produce the Multimedia results.

[1] http://lucene.apache.org/solr/

[2] http://lucene.apache.org/java/docs/index.html

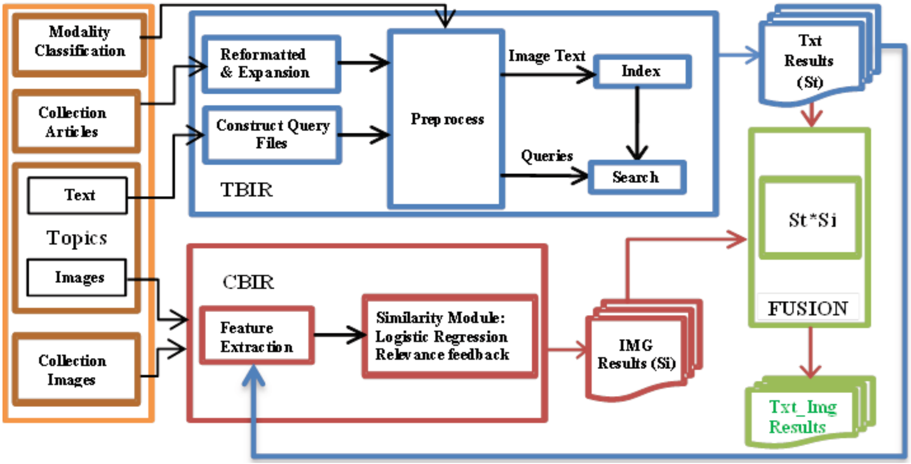


Fig. 2. System Overview showing Textual, Visual and Fusion Subsystems

The work of the **CBIR subsystem** is based on two main stages: Extraction of the low-level visual features of the images, and the calculation of the similarity (Si) for each image of the collection to the image examples given by a query.

1. **Extraction of low-level features:** The first step in the CBIR system is to extract the visual features for all the images in the database, as well as for the example images given in each question. The features calculated by the group are:

- Color information: Color information has been extracted by calculating both local and overall histograms of the images. Overall histograms have been calculated using 10x3 bins on the HS (Hue Saturation) color system. Meanwhile, local histograms have been calculated by dividing the images into four fragments of the same size. A bi-dimensional HS histogram with 12x4 bins is computed for each patch. Therefore, a feature vector of 222 components represents the color information of the image.
- Texture information: This information is embodied as the granulometric distribution function. A granulometry is defined from the morphological opening of the texture using a convex and compact subset containing the origin as structuring element [11]. In our case we have used a horizontal and a vertical segment as the structuring elements, being 60 components in total for both structuring elements. And the Spatial Size Distribution that is another morphological operation defined in [1] using a horizontal segment as structuring element, being 10 components.

2. **Similarity Module:** The similarity module uses our own logistic regression relevance feedback algorithm [9]. The algorithm calculates the probability of an image belonging to a set of those images sought by the query and models the *logit* of this probability as the output of a generalized linear model whose inputs are

the visual low-level image features. The algorithm needs examples and counter-examples (positive and negative images). The positive images are the example images of the topic given by the organization (from one to three in the Medical2011 edition). As the Medical2011 does not provide any set of non-relevant images, the M counter-examples are obtained by applying a procedure, which chooses J random images from the whole database (without images in the pre-filtered textual list). The Euclidean distance ranks these J images, and the latest M images are taken as negative examples.

In the following some details are given about the way the logistic regression relevance feedback algorithm works. Let us consider the (random) variable Y giving a user evaluation in which Y=1 means that the image is positively evaluated and Y=0 means a negative evaluation. Each image in the database has been previously described by using low-level features in such a way that the *j-th* image has the *k*-dimensional feature vector x_j associated. Our data consist of (x_j,y_j), with $j=1.....n$, where n is the total number of images, x_j is the feature vector and y_j the image evaluation (1=positive and 0=negative). The image feature vector x is known for any image and we try to predict the associated value of Y. In this work we have used a logistic regression where P(Y=1|x) i.e. the probability that Y=1 (the image is positively evaluated) given that the feature vector x is related to the systematic part of the model (a linear combination of the feature vector) by means of the *logit* function. For a binary response variable Y and p explanatory variables $x_1.....x_p$ the model for $\pi(x)=P(Y=1|x)$ at values $x=(x_1.....x_p)$ of predictors is logit $[\pi(x)]=\alpha+\beta_1x_1+...+\beta_px_p$. Where logit $[\pi(x)]=\ln(\pi(x)/(1-\pi(x)))$. The model parameters are obtained by maximizing the likelihood function given by:

$$l(\beta)=\prod^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{1}$$

The maximum likelihood estimator (MLE) of the parameter vector β is calculated by using an iterative method. We have a major difficulty when having to adjust an overall regression model in which we take the whole set of variables into account because the number of selected images (the number of positive plus negative images, k) is typically smaller than the number of characteristics (k < p). In order to solve this problem our proposal is to adjust different smaller regression models.

The **fusion subsystem** is in charge of merging the two score result lists from the TBIR and the CBIR subsystem. In the present work we use the product fusion algorithm (Si*St). The two results lists are fused together to combine the relevance scores of both textual and visually retrieved images (St and Si). Both subsystems will have the same importance for the resulting list. The normalization is an important issue to take into account when working with multimodal information. We propose that different value modalities were within the same range of between 0 and 1, but the normalization is independent for each one. The idea is that each module obtains its score between the predefined ranges in order not to normalize to the maximum for each independent query. The objective of this idea is that each module relies on its own confidence. The CBIR subsystem meets this condition so that the images score (Si) is the probability that a given image belongs to a certain set of images. A Si score would be 1 if the CBIR were completely sure that the image belongs to the set, so that the probabilities obtained are close to 1 if they are very similar to the given examples.

To obtain the TBIR similarity score (St) of a document for a specific query, *Lucene* does not exactly calculate the cosine measure (as supposed in the VSM approach) for reasons of usability [2]. The original cosine function normalizes the dot product of the weighted vectors ($V(q)\cdot V(d)$) by dividing it by the Euclidean norm of these vectors ($|V(q)|$ and $|V(d)|$), normalizing them into a unit vector. *Lucene* uses a different document length normalization for *V(d)*, which normalizes it into a vector equal to or larger than the unit vector. Instead of calculating the Euclidean norm for each document, which would have a very high computational cost, *Lucene* uses the document length, which is computed when the document is added to the index in accordance with the number of tokens in the document. For efficient score computation, the query Euclidean norm ($|V(q)|$) is computed when search starts, as it is independent of the document being scored. Normalizing the query vector *V(q)* provides comparability (to a certain extent) of two or more queries. That is why the normalization is calculated independently for CBIR and TBIR.

4 Experiments Regarding Query Modality Classes

The main goal of this paper is the experimentation carried out to show how and when results can be improved using modality-related challenges. The comparison between these results shows that is query-dependent. The textual approaches studied are:

- Query reformulation by removing "*domain stopwords*" and
- Inclusion of the modality classification in the search process.

For this purpose at different experiments (runs), during preprocessing, the identified *domain stopwords* terms related to the modality like, "view" or "images" are deleted. Also in the preprocess step, the terms likely to be representatives of a AND operator (e.g. with) are converted to AND. After, the modality class identification task is manually performed for each query (30), and it is added to the original query, when possible. For example, in query ***x-ray images of a tibia with a fracture***, the x-ray modality class information can be identified but in query ***photographs of benign or malignant skin lesions*** it is impossible to identify an associated modality. Note that the modality is explicit only in 13 of the 30 queries.

The inclusion of the modality classification in every query is carried out as follows:

1. Manual identification of the modality classification that is expected for the results from the query.
2. Deletion of stopwords as well as the *domain stopwords* (these are the experiments without modality expansions are denoted as runIC1 and runIC2).

3. Query expansion with the information of the tag field of images. The field can be aggregated (with OR) to the original query or used like a filter (with AND). Experiments with modality expansion are denoted as runIC3 and runIC4. For example:

   Original query: chest CT images with emphysema[3].
   OR aggregation: (tags:CT)^0.7 OR (chest CT AND emphysema)^0.3
   AND aggregation: (tags:CT) AND (chest CT AND emphysema)

Table 1 shows the experiments (runs) based on the textual retrieval sent to the 2011 campaign (i.e. regardless the results of the CBIR subsystem). The evaluation is based on three different measurements: MAP (Mean Average Precision), P@10 (precision at first 10 images retrieved) and the recall (number of relevant image retrieved).

Table 1. ImageCLEF 2011 Text-based Results.

Name	Type of query	MAP	P@10	Recall
RunIC1	Modality information NOT expanded	0.2158	0.3533	0.5352
RunIC2	Modality information expanded	0.2125	0.3867	0.4404
RunIC3	With search against tag field (OR)	0.1309	0.3433	0.2183
RunIC4	With search against tag field(AND)	0.1270	0.3100	0.2519
Best Textual	Run	0.2172	0.3467	0.5693

The best run for textual modality was the Laberinto_CTC run with a MAP value of 0.2172 [8] that is close to our best result (0.2158 for RunIC1). In our text-based approaches the inclusion of modality classification significantly worsen the overall results: runs with modality expansion (runs IC3 and IC4 have lower MAP and recall values than runs IC1 and IC2 that do not use modality expansion).

Given these results, it was decided to investigate the behavior of the modality expansion in more detail. Based on this, the Road Map with different experiments performed is presented in the following. The detailed information of the experiments is:

- **Run 1:** This is the baseline (the original query without modality). The search is carried out against the caption field of the images. For example,
  Original query: chest CT images with emphysema.
  Run 1: chest AND emphysema.
- **Run 2:** In this case the modality information is not removed or processed in any way (as was in RunIC1). The search is carried out against the caption field of the images. For the example,
  Original query: chest CT images with emphysema.
  Run 2: (chest CT) AND emphysema.
- **Run 3:** The modality information is expanded with the label of the modality established in the ImageCLEF collection as in Run IC2 (see the description of the collection at Section 2). The search is performed against the caption of images. Example:

[3] When the operator between terms is not specified, it is understood an OR operator.

Modality: CT **Molality Label:** computed tomography
Original query: chest CT images with emphysema.
Run 3: (chest (CT OR "computed tomography")) AND emphysema

- **Run 4:** Similar to Run 3, but the modality in the original query is removed. The search is carried out against the tags field of the images in the collection, as well as in the caption of images in Run3. Example:

  Original query: chest CT images with emphysema.
  Run 4: (tags: CT)^0.7 OR (chest AND emphysema)^0.3

- **Run 5:** Like Run 4 but using the tags field such as filter instead of aggregation. For example,

  Original query: chest CT images with emphysema.
  Run 5: (tags: CT) AND (chest AND emphysema).

Table 2 includes the quantitative results for the new experiments analysis in terms of MAP, P@10 and recall.

Table 2. Results of the new experiments proposed.

Name	Experiment description	MAP	P@10	Recall
Run 1	Baseline deleting modality	0.2003	0.3333	0.5186
Run 2	Baseline	0.2158	0.3533	0.5352
Run 3	Modality + Modality Information	0.2125	0.3867	0.4404
Run 4	Original query without modality OR Modality query against tag field	0.1883	0.3667	0.4172
Run 5	Original query without modality AND Modality query against tag field	0.1932	0.3793	0.3824

Table 2 shows that modality information provides better results than not using it (Run2 outperforms Run1). Nevertheless, the runs that use the modality information to expand the query (Runs 3 to 5) do not outperform the baseline Run2 in terms of MAP and recall, but they do in terms of P@10. The modality information makes the query be more restrictive so for the queries that this information is not really useful makes that less relevant images do not be retrieved (decreasing the recall and the MAP). This restrictiveness makes also be more efficient at first position retrievals (increase P@10). It is also importance to notice the new experiments using modality information to expand the queries (Runs 4 and 5) improve significantly our ImageCLEF runs with modality expansion (RunIC3 and RunIC4) in terms of MAP, P@10 and recall.

For a deeper analysis, a query-by-query analysis was carried out taking into account only the queries in which the modality can be identified (13 of 30 shown in Table 3) and, consequently, the modification of the queries can have an influence. Table 3 shows a global analysis of the queries that includes modality classification. Figures 3, 4 and 5 show the qualitative results for the query-by-query analysis for runs 1, 2 and 5, respectively. The recall of each query is shown in bars, according to the scale on the left, and the MAP with points and squares according to the scale on the right. The squares denote the queries with modality information and the points the rest of the queries. Figure 3 and 4 show containing results of queries with modality information (squares in figure) in Run 2 are better than in Run 1.

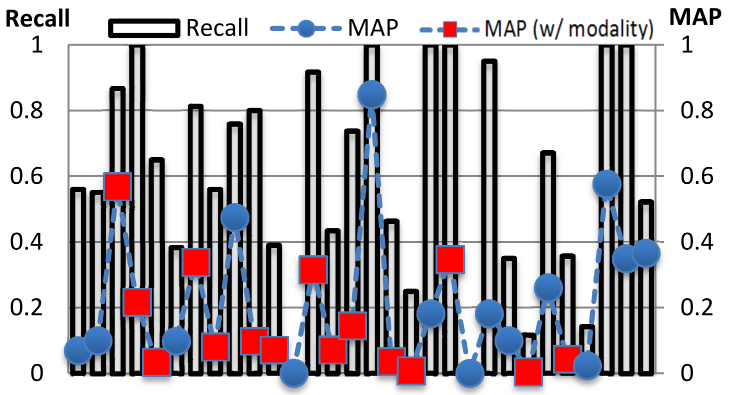


Fig. 3. Results Query by query (run1)

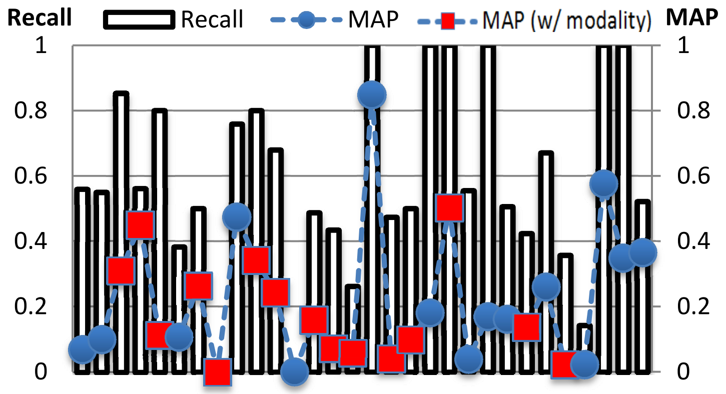


Fig. 4. Results Query by Query (run 2)

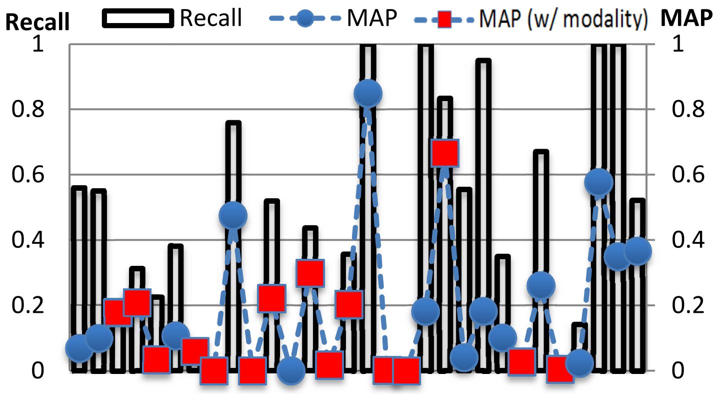


Fig. 5. Results Query by Query (Run 5)

Table 3. Qualitative results: Query comparison respect baseline without modality (Run1)

#	Query text	Run 2 (Baseline)	Run 5 (original query without modality + modality)
3	Doppler ultrasound images (colored)	MAP decrease, given the increment in docs retrieved.	Large decrease in MAP and relevant results.
4	chest CT images with emphysema	MAP increase but fewer relevant results	Similar MAP but much fewer relevant results
5	sagittal views of head MRI images	MAP and relevant results increase	MAP and relevant results decrease
7	x-ray images of a tibia with a fracture	MAP and number of relevant results decrease	Large decrease in MAP and in number of relevant results
8	x-ray images of a hip joint with prosthesis	MAP decrease	MAP decrease
10	medial meniscus MRI	MAP increase	MAP decrease. No documents retrieved
11	abdominal CT images showing liver blood vessels	MAP and number of relevant results increase	MAP and number of relevant results increase
13	all x-ray images containing one or more fractures	Decrease in MAP and number of relevant results	Similar results but increase in MAP
14	Angiograms containing the aorta	Similar result	Large decrease of number of relevant results and MAP
15	CT images with a brain infarction	Decrease in MAP and number of relevant results	MAP increase but produced by the decrease of total results
17	microscopic pathology images of the kidney	Similar results	Large decrease in number of retrieved results and consequently MAP too
18	fetal MRI	MAP increase	Large decrease in number of retrieved results and MAP
20	CT liver abscess	Large increase in MAP	Large increase in MAP
24	CT or x-ray images showing the heart	Large increase in MAP	Increase in MAP
26	microscopic images of tissue from the cerebellum	Similar or worse result	Worse result

Looking in detail Table 3, for the queries that do not improve the results, some interesting details can be extracted:

- Taking only Run 2 into account some queries retrieved worse or similar results than the baseline run in terms of MAP and also in the number of relevant documents retrieved (marked in bold in Table 3). These queries (#7, #13, #14 and #15) include implicitly the modality in the query, even if there were not modality explicitly. In these queries, the expansion includes redundant information (for instance in query #13, fractures are typical of x-ray images).
- Run 5 shows that, in some cases, there are some queries in which modality expansion is not useful for expansion, due to the nature of the query. An example of this is “microscopic” modality, with label EM, (queries #17 and #26). This is perhaps due to the classification is not correct, because it is done in an automatic way without supervision, or maybe that our approach doesn’t work with this kind of queries.

Table 4 shows the textual, the image and the merged results for runs 1 to 5 with the aim to evaluate the improvement of merging multimodal information. It can be observed that the visual runs have lower MAP than textual results due to the fact that low-level features alone do not sufficiently capture the meaning of the topics, as it is already known at literature. It can be also observed that the mixed experiments from runs 1 to 3 outperform the textual ones. Nevertheless, the fusion results for runs 4 and 5 do not outperform their textual baselines. We think this is due to the fact that these runs have low recall values (0.3824 at run 5) that means that 62% of relevant images have not at the pre-filter textual list, and are not then evaluated by the image module. For improving these fusion results we should improve first the textual baseline results.

Table 4. Fusion Results using the product of Textual and Visual Scores (St*Si)

Name	Text		Image		Text-Image	
	MAP	P@10	MAP	P@10	MAP	P@10
Run 1	0.2003	0.3333	0.1231	0.2667	0.2110	0.3567
Run 2	0.2158	0.3533	0.1093	0.1800	0.2297	0.4000
Run 3	0.2125	0.3867	0.1061	0.2633	0.2226	0.4067
Run 4	0.1883	0.3667	0.1165	0.2200	0.1754	0.3333
Run 5	0.1932	0.3793	0.1412	0.2828	0.1874	0.3759

5 Conclusions and Future Work

The query expansion with modality classes improves the results in overall and in query-by-query analysis, but this improvement is dependent on the query type. So, to carry out the expansion, the type of query should be identified before the expansion (as can be also viewed in other Information Retrieval Tasks as Question Answering).

Another aspect that affects the results is the accurate modality classification of the images as well as a better textual representation in the captions. When taxonomy is used in order to classify a collection, the retrieval results are not very accurate if taxonomy is not well represented in the collection. As previously discussed, for some cases, the modality labels are not as precise as it should be (e.g. EM label for microscope images), or even incorrect. In fact, the organization has changed the classification for the Medical Retrieval Task this year, expanding the number of categories from 18 to 31 [10].

One point to remark is that, as in all of our previous work, the fusion of textual and visual results improves the results separately, but only when the recall of textual filter is high enough.

Motivated by the conclusions of this work, we intend to work in the automatic detection of the type of the query to adjust the expansion process by applying this information. It is expected that a more accurate expansion can improve overall results instead of only those that have a modality associated. Our aim is also to apply this methodology to other domain-dependent repositories in addition to the medical one. The multilingualism has not been used in this work due to the captions and queries in French and German are simply direct translation from English. We assumed that these translations were correct; however, it could be interesting to study the effect of using these multilingual descriptions.

Acknowledgments. This work has been partially supported for Regional Government of Madrid under Research Network MA2VIRMR (S2009/TIC-1542), for Spanish Government by project BUSCAMEDIA (CEN-20091026), by project HOLOPEDIA (TIN 2010-21128-C02) and by project MCYT TEC2009-12980.

References

1. Ayala, G., Domingo, J.: Spatial Size Distributions. Applications to Shape and Texture Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1430–1442 (2011)
2. Bedrick, S., Kalpathy-Cramer, J.: Improving Retrieval Using External Annotations: OHSU at ImageCLEF 2010. In: Working Notes for the CLEF 2010 (2010)
3. Benavent, J., Benavent, X., de Ves, E., Granados, R., García-Serrano, A.: Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information approaches. In: Working Notes for the CLEF 2010 (2010)
4. Castellanos, A., Benavent, X., Benavent, J., García-Serrano, A.: UNED-UV at Medical Retrieval Task of ImageCLEF 2011. In: Working Notes of the CLEF 2011 (2011)
5. Chevallet, J., Lim, J.: Using Ontology Dimensions and Negative Expansion to solve Precise Queries in CLEF Medical Task. In: Working Notes of the CLEF 2005 (2005)
6. Clinchant, S., Csurka, G., Ah-Pine, J., Jacquet, G., Perronin, F., Sánchez, J., Minoukadeh, K.: XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classification Ad-hoc Retrieval Tasks of ImageCLEF 2010. In: Working Notes of the CLEF 2010 (2010)
7. Granados, R., Benavent, J., Benavent, X., de Ves, E., García-Serrano, A.: Multimodal information approaches for the Wikipedia collection at ImageCLEF 2011. In: Working Notes of the CLEF 2011 (2011)
8. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I.: Garcia Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of the CLEF 2011 (2011)
9. Leon, T., Zuccarello, P., Ayala, G., de Ves, E., Domingo, J.: Applying logistic regression to relevance feedback in image retrieval systems. Pattern Recognition 40, 2621–2632 (2007)
10. Müller, H.: Garcia Seco de Herrera, A., Kalpathy-Cramer, J., Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: Working Notes of the CLEF 2012 (2012)
11. Soille, P.: Morphological Image Analysis: Principles and Applications. Springer, Berlin (2003)
12. Tirilly, P., Lu, K., Mu, X., Zhao, T., Cao, Y.: On modality classification and its use in text-based image retrieval in medical databases. In: 9th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 109–114 (2011)
13. Torjmen, M., Pinel-Sauvagnat, K., Boughanem, M.: Methods for Combining Content-Based and Textual-Based Approaches in Medical Image Retrieval. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 691–695. Springer, Heidelberg (2009)
14. Villena-Román, J., Lana-Serrano, S., González-Cristóbal, J.-C.: MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 593–596. Springer, Heidelberg (2008)

Author Index